JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7759

REAL-TIME VIDEO-BASED VISITOR COUNTING FOR SMART TOURISM DESTINATIONS USING YOLOV11 AND BYTETRACK

Feriantano Sundang Pranata[1*]; Arif Adrian[1]; Khairani Saladin[1]

Departemen Pariwisata[1]
Universitas Negeri Padang, Padang, Indonesia[1]
https://www.unp.ac.id[1]
feriantano@unp.ac.id*, arif.adrian@fpp.unp.ac.id, khairanisaladin@fpp.unp.ac.id

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *Accurate and real-time visitor data are needed to support smart tourism management. However, conventional counting methods still have limitations in dynamic outdoor tourism environments. This study develops and evaluates a real-time video-based visitor counting system by integrating YOLOv11 for person detection and ByteTrack for multi-object tracking. This approach extends visitor counting evaluation to uncontrolled open-air tourist destinations, where lighting variation, background complexity, visitor movement, and crowd density may affect detection and tracking performance. The system was evaluated using nine Full HD videos from five tourist destinations in West Sumatra, recorded under daylight and afternoon conditions with low to medium visitor densities. The YOLOv11–ByteTrack system achieved an average counting accuracy of 84.02%, MAE of 7.22 visitors per video, MAPE of 15.98%, and an average processing speed of 36.23 FPS. The average accuracy exceeded those of YOLOv3 and YOLOv8, which achieved 75.71% and 77.15%, respectively. These findings suggest that YOLOv11–ByteTrack has practical potential as a real-time visitor counting approach in smart tourism management, particularly for monitoring visitor flows, assessing site capacity, controlling visitor density, and supporting data-driven infrastructure planning.*

Keywords: *ByteTrack, Multi-Object Tracking, Smart Tourism, Visitor Counting, YOLOv11.*

Intisari— *Data pengunjung yang akurat dan real-time dibutuhkan untuk mendukung pengelolaan wisata cerdas. Namun, metode penghitungan konvensional masih memiliki keterbatasan pada lingkungan wisata luar ruang yang dinamis. Penelitian ini mengembangkan dan mengevaluasi sistem penghitungan pengunjung berbasis video real-time dengan mengintegrasikan YOLOv11 untuk deteksi manusia dan ByteTrack untuk pelacakan multi-objek. Pendekatan ini memperluas evaluasi penghitungan pengunjung ke destinasi wisata luar ruang yang tidak terkontrol, dengan variasi pencahayaan, kompleksitas latar belakang, pergerakan pengunjung, dan kepadatan kerumunan yang dapat memengaruhi kinerja deteksi dan pelacakan. Sistem dievaluasi menggunakan sembilan video Full HD dari lima destinasi wisata di Sumatera Barat yang direkam pada kondisi siang dan sore hari dengan kepadatan pengunjung rendah hingga sedang. Sistem YOLOv11–ByteTrack menghasilkan akurasi penghitungan rata-rata 84,02%, MAE 7,22 pengunjung per video, MAPE 15,98%, dan kecepatan pemrosesan rata-rata 36,23 FPS. Akurasi tersebut lebih tinggi dibandingkan YOLOv3 dan YOLOv8, yang masing-masing mencapai 75,71% dan 77,15%. Temuan ini menunjukkan potensi praktis penggunaan YOLOv11–ByteTrack sebagai pendekatan penghitungan pengunjung secara real-time dalam pengelolaan wisata cerdas, khususnya untuk pemantauan arus pengunjung, penilaian kapasitas, pengendalian kepadatan, dan perencanaan infrastruktur berbasis data.*

Kata Kunci: *ByteTrack, pelacakan multi-objek, wisata cerdas, penghitungan pengunjung, YOLOv11.*

INTRODUCTION

The tourism sector relies heavily on effective visitor data management to support infrastructure planning, capacity control, and evidence-based decision-making. In the context of smart tourism, digital technologies play a key role in enhancing destination management efficiency and improving the tourist experience [1], [2], [3], [4]. Data-driven management enables decision-makers to understand tourist behavior patterns, predict visitation surges, and adapt strategies to on-site conditions [5], [6], [7], [8]. Therefore, accurate and up-to-date visitor data is essential for developing sustainable, demand-responsive tourist destinations [9].

Despite the advantages of digital tools, many tourist destinations continue to rely on conventional counting approaches and monitoring systems [10]. These methods are prone to errors and may reduce data reliability, particularly in busy environments [11]. Moreover, they create operational inefficiencies, particularly at high-traffic sites, leading to long queues and disrupted visitor flow. These inefficiencies adversely affect the quality of decisions regarding capacity allocation, facility management, and safety measures at tourist destinations [12], [13], [14].

Although conventional monitoring technologies and sensor-based approaches have been used for visitor counting [15], open and dynamic outdoor settings introduce additional challenges for visitor counting systems [16], [17], which are typical of natural and cultural tourism destinations. Furthermore, many conventional systems fail to provide real-time data, a critical issue in today's digital age. Delayed or inaccurate data complicates decision-making, making it difficult for managers to adjust policies in real time, such as reallocating staff or managing parking during peak hours [18], [19]. Additionally, the lack of historical data for analyzing visitation trends impedes the ability to plan promotions and implement evidence-based management strategies [7], [8].

Recent studies have explored imagery- and camera-based approaches as alternatives to traditional visitor counting methods [15], [16], [20], [21]. For instance, camera traps combined with machine-learning-based computer vision have been used to support visitor monitoring in outdoor recreation areas [16]. Other studies have applied YOLO-based and deep learning-based models for real-time crowd detection, public-space monitoring, and occupancy estimation in structured environments [20], [22], [23]. These studies indicate that computer vision can provide a non-intrusive and scalable solution for people counting and crowd monitoring. In the tourism domain, computer vision has also been recognized as a promising component of smart destination systems, particularly for improving monitoring, service management, and visitor analytics [17].

Despite these developments, several limitations remain in the existing literature. First, many previous studies have been conducted in indoor, semi-controlled, or general surveillance environments, where lighting conditions, camera viewpoints, background characteristics, and movement patterns are relatively more stable [20], [22], [23]. Second, studies that specifically evaluate real-time visitor counting in open-air tourist destinations remain limited, even though such environments are characterized by fluctuating illumination, complex backgrounds, partial occlusion, and varying visitor densities [17], [21]. Third, comparative evaluations of recent YOLO-based detection models integrated with tracking algorithms for tourism-specific visitor counting are still insufficient. As a result, the practical performance of a detection–tracking pipeline in uncontrolled outdoor tourism contexts has not been fully established.

To address this gap, this study develops and evaluates a real-time video-based visitor counting system for open-air tourist destinations by integrating YOLOv11 and ByteTrack. YOLOv11 detects visitors as person objects in video frames, while ByteTrack maintains identity consistency across frames to reduce duplicate counting. The system is evaluated using video data from five tourist destinations in West Sumatra under varying lighting conditions and crowd densities, with performance assessed through counting accuracy, error-based metrics, and processing speed.

This study contributes to the literature and practice of smart tourism in four ways. First, it presents a unified detection–tracking–counting pipeline that integrates YOLOv11 and ByteTrack for real-time visitor counting. Second, it evaluates the system in open-air tourist destinations, a context that remains less frequently examined than indoor or controlled monitoring environments. Third, it compares the counting performance of YOLOv11 with YOLOv3 and YOLOv8 under the same dataset and evaluation setting. Fourth, it provides quantitative evidence on the feasibility of using a real-time video-based counting system to support visitor flow monitoring, capacity assessment, and data-driven tourism infrastructure planning.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7759

MATERIALS AND METHODS

This study adopts an applied experimental design focusing on the development and quantitative evaluation of a video-based visitor counting system using computer vision and deep learning approaches. The main model employed is YOLOv11 as the object detector, integrated with the ByteTrack tracking algorithm to maintain identity consistency across video frames. The system development process was conducted iteratively through requirements analysis, system architecture design, implementation of a pre-trained YOLOv11 detection model, integration with ByteTrack, and performance evaluation under operational conditions at tourist destinations.

The system was designed to automatically detect, track, and count visitors from video recordings in real time. Conceptually, the workflow consists of video preprocessing, object detection using YOLOv11, object tracking with ByteTrack, and visitor counting using a virtual counting-line strategy. These stages form an integrated detection-tracking-counting framework that enables the system to estimate visitor counts in open-air tourism environments characterized by dynamic lighting, diverse backgrounds, and varying crowd densities.

Dataset Description

The dataset used in this study consists of nine videos recorded at five tourist destinations in West Sumatra: Puncak Gado-gado (PG), Gunung Padang (GP), Pantai Air Manis (AMB), Pantai Carocok (CB), and Pantai Gandoriah (GB). Each video has a duration of between eight and twelve minutes, with an original resolution of 1920 x 1080 pixels (Full HD). The recording conditions varied in terms of illumination, visitor density, environmental background, and camera viewpoint. These variations were deliberately included to assess the robustness of the proposed system in outdoor tourism environments with fluctuating illumination, visitor movement, and crowd density.

The selection of recording locations was based on several criteria. The selected destinations are popular and accessible, enabling the capture of different visitor volumes under realistic tourism conditions. The dataset includes low- and medium-density scenarios to evaluate the system across different crowd levels. Another consideration was the feasibility of stable camera installation to minimize instability, motion blur, and excessive viewpoint changes. The videos were recorded under daylight and afternoon conditions to represent common illumination variation in real outdoor environments.

Ground-truth data were obtained through manual counting of visitors by a single observer for each recording. The observer counted individuals visible in the video and recorded the number of visitors based on visible presence. This manual count served as the reference for validating the automatic counting system. However, potential bias may occur because only one observer was involved, particularly in cases of partial occlusion, overlapping visitors, or ambiguous object boundaries. Therefore, the ground-truth data should be interpreted as a practical reference for evaluation rather than an error-free measurement. A detailed description of the dataset is provided in Table 1, where various attributes of the dataset are outlined. Sample frames extracted from the dataset are presented in Figure 1, which illustrates the visual characteristics and variability of the input data used for analysis.

Table 1. Dataset Overview

No	Name	Frames (Duration)	Ground-truth	Condition
1	PG-01	15659 (08:41)	45	Afternoon, medium density
2	PG-02	16403 (09:06)	33	Afternoon, medium density
3	GP-01	22149 (12:18)	85	Daylight, medium density
4	GP-02	16699 (09:16)	56	Daylight, medium density
5	AMB-01	14813 (08:13)	40	Daylight, low density
6	CB-01	15253 (08:45)	28	Daylight, low density
7	CB-02	14961 (08:35)	41	Daylight, low density
8	GB-01	15602 (08:40)	38	Daylight, medium density
9	GB-02	14917 (08:17)	20	Daylight, medium density

Source: (Research Results, 2025)

Source: (Research Results, 2025)

Figure 1. Sample Frames from the Dataset Showing Variations in Lighting, Crowd Density, Background, and Viewpoint

Experimental Setup

All experiments were conducted on a computer system equipped with Windows 11 Pro, an Intel Core i5-11400H CPU running at 2.70 GHz,

16 GB of RAM, and an NVIDIA RTX 3050 Ti GPU with 4 GB of VRAM. Video data were captured using a 108-megapixel smartphone camera to obtain high-resolution input suitable for diverse environmental conditions. The implementation was developed using Python 3.11 with supporting libraries including PyTorch, OpenCV, ByteTrack, and Flask. This computational environment was selected to represent a mid-range processing platform capable of performing real-time inference, thereby demonstrating the practical feasibility of the proposed system for smart tourism applications.

The object detection module used the YOLOv11 model provided by the Ultralytics framework. The model was used in an off-the-shelf configuration with pre-trained weights from the COCO dataset, and only the person class (class ID = 0) was detected. Fine-tuning was not performed because this study aimed to evaluate the baseline feasibility of a readily deployable detection model for real-time visitor counting in outdoor tourism environments. This design choice also reflects practical deployment conditions in which locally annotated tourism datasets may not always be available.

The absence of fine-tuning may affect detection accuracy because the visual characteristics of local tourism scenes, such as non-standard camera viewpoints, complex outdoor backgrounds, partial occlusion, and overlapping visitors, are not specifically represented in the training data. Therefore, the reported performance should be interpreted as the baseline capability of a pre-trained YOLOv11 model integrated with ByteTrack, rather than as the maximum achievable performance after local model adaptation.

For model inference, the YOLOv11 input image size was set to 640 x 640 pixels. This inference size was selected to balance processing speed and detection accuracy, enabling real-time processing while maintaining sufficient spatial information for human detection. The detection confidence threshold was set to 0.25, while the Intersection over Union (IoU) threshold was set to 0.45 to balance object separation and false detection reduction. To improve reproducibility, the main experimental parameters used in the detection, tracking, and counting stages are summarized in Table 2. Parameters not explicitly modified were retained at their default values in the Ultralytics YOLOv11 and ByteTrack implementations.

Table 2. Experimental Parameter Configuration

Component	Parameter	Value	Description
YOLOv11	Pre-trained weights	COCO	Off-the-shelf model without fine-tuning
YOLOv11	Detected class	Person / class ID 0	Only human objects were detected
YOLOv11	Confidence threshold	0.25	Minimum confidence for valid detection
YOLOv11	IoU threshold	0.45	Threshold for non-maximum suppression
YOLOv11	Inference image size	640 x 640	Input size for model inference
ByteTrack	track_high_thresh	0.25	Threshold for high-confidence association
ByteTrack	track_low_thresh	0.10	Threshold for low-confidence association
ByteTrack	new_track_thresh	0.25	Threshold for initializing new tracks
ByteTrack	track_buffer	30 frames	Number of frames to retain lost tracks
ByteTrack	match_thresh	0.80	Matching threshold for track association
ByteTrack	fuse_score	Enabled	Combines detection confidence and association score
Counting	Counting direction	One-directional	Reverse crossings were ignored
Counting	Counting trigger	Line crossing	Count increased when object centroid crossed the line

Source: (Research Results, 2025)

System Architecture

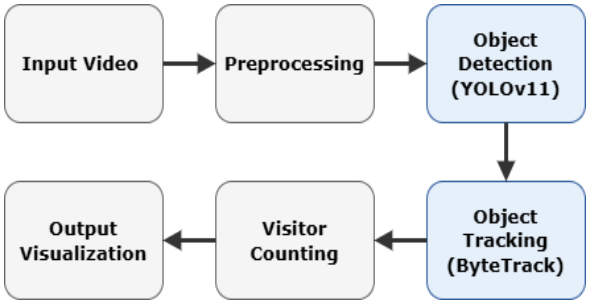
The proposed system architecture consists of four core processing modules, namely video preprocessing, object detection, object tracking, and visitor counting, supported by input video acquisition and output visualization components. In the preprocessing stage, the input videos with an original resolution of 1920 x 1080 pixels were decomposed into individual frames using the OpenCV library. For application-level processing and visualization, the frames were resized to 1280 x 720 pixels to reduce computational load while preserving sufficient visual detail. During the detection stage, YOLOv11 performed inference using an input image size of 640 x 640 pixels, as specified in the experimental configuration. Thus, the 1280 x 720 resolution was used for frame handling and output visualization, whereas 640 x 640 was used as the model inference size.

The object detection stage identifies visitors as persons in each frame and produces bounding box coordinates and confidence scores. These detection outputs are subsequently passed to ByteTrack for multi-object tracking. ByteTrack maintains object identity continuity across consecutive frames by associating high-confidence and low-confidence detections, thereby reducing trajectory loss during partial occlusion or temporary detection failure.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7759

The visitor counting stage uses a configurable virtual counting-line mechanism. The counting line was placed at strategic entry or flow points within the camera field of view according to the spatial layout of each tourist destination. The camera was mounted on a fixed tripod at an approximate height of 2.5 to 3 meters above ground level with a downward diagonal angle of 30-45 degrees. The camera remained static throughout the recording process to support consistent frame-to-frame analysis and reliable trajectory tracking. The overall workflow of the proposed system is illustrated in Figure 2.



Source: (Research Results, 2025)
Figure 2. Workflow of the Proposed Visitor Counting System Integrating YOLOv11 Detection, Bytetrack Tracking, and Line-Crossing-Based Counting

For each detected and tracked visitor, the centroid of the bounding box was calculated to determine the object position across consecutive frames. Given a bounding box represented by its top-left and bottom-right coordinates, the centroid point was computed using Equation (1).

$$x_c = \frac{(x_1 + x_2)}{2}, \quad y_c = \frac{(y_1 + y_2)}{2} \qquad (1)$$

where (x_1, y_1) and (x_2, y_2) denote the top-left and bottom-right coordinates of the bounding box, respectively, while (x_c, y_c) represents the centroid of the tracked object.

The counting mechanism was implemented using axis-aligned virtual lines placed at predefined positions in the resized frame. A vertical counting line was represented by a fixed x-coordinate L_x, while a horizontal counting line was represented by a fixed y-coordinate L_y. Depending on the selected monitoring direction, the system evaluated whether the centroid of the same tracked object moved across the selected threshold between two consecutive frames.

Table 3. Virtual Counting-Line Configuration Used in the Implementation

Line type	Variable	Position	Movement
Left vertical	L_{left}	0.10W	Right-to-left
Right vertical	L_{right}	0.90W	Left-to-right
Bottom horizontal	L_{bottom}	0.90H	Top-to-bottom
Center horizontal	L_{center}	0.70H	Bottom-to-top

Source: (Research Results, 2025)

In Table 3, W and H denote the width and height of the resized frame, respectively. In this implementation, W = 1280 pixels and H = 720 pixels. A crossing event was determined by comparing the centroid position of the same tracked object between two consecutive frames. The crossing status was formulated as Equation (2).

$$Cross_t = I\{C_v, t \text{ or } C_h, t\} \qquad (2)$$

where $Cross_t$ denotes the crossing status at frame t, and $I\{\cdot\}$ is an indicator function that returns 1 when the condition is satisfied and 0 otherwise. C_v, t represents a vertical-line crossing in the predefined direction, while C_h, t represents a horizontal-line crossing in the predefined direction. For combined direction settings, such as bottom_left or bottom_right, a crossing event was recorded when either the corresponding vertical-line condition or horizontal-line condition was satisfied.

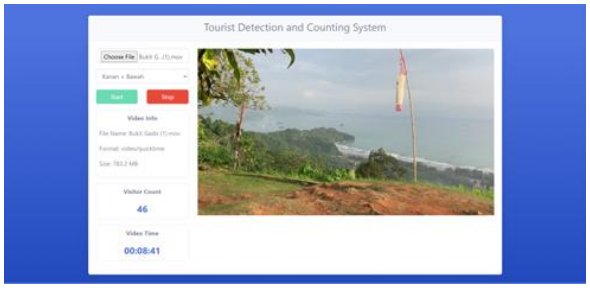
The visitor counter was incremented only when a tracked object crossed the selected counting line in the predefined direction and its identity had not been counted before, as shown in Equation (3).

$$Count = Count + 1, if\ Cross_t = 1\ and\ ID \notin counted_{IDs} \qquad (3)$$

where $counted_{IDs}$ denotes the set of tracking identities that had already been counted. In the implementation, this mechanism was represented by a per-ID counting status to ensure that each tracked visitor was counted only once after crossing the selected virtual line. Reverse crossings were ignored based on the selected counting direction to reduce double counting in the evaluated scenarios.

Implementation Details

The system was implemented as a web-based application that enables users to upload videos, run the visitor detection process, and obtain analysis results through an interactive interface. This design allows the detection workflow to be performed efficiently through a standard web browser. Figure 3 shows the web-based system interface.



Source: (Research Results, 2025)
Figure 3. User Interface of the Web-Based Visitor Counting System

As shown in Figure 3, the system interface displays processed videos annotated with bounding boxes indicating detected visitors in each frame. The interface also provides real-time information on the number of detected visitors during video playback. In addition, the system reports the processing time required for each video along with the total number of visitors counted. These features enable users to observe both the visual detection results and the system performance in a clear and comprehensive manner.

Evaluation

The system was evaluated in terms of counting effectiveness, functional reliability, and computational efficiency within the operational context of tourist destinations. The evaluation focused on three aspects: visitor counting accuracy, system functionality, and system performance. This evaluation design ensured that the system was assessed not only based on counting accuracy, but also on operational stability and real-time processing capability.

1. Counting Accuracy Evaluation

Counting accuracy was evaluated by comparing the visitor counts estimated by the system with the ground-truth counts obtained through manual observation at the same locations. The accuracy metric was computed using the relative error with respect to the ground-truth count, as defined in Equation (4).

$$A = \left(1 - \frac{|C_s - C_m|}{C_m}\right) \times 100\% \quad (4)$$

where C_s denotes the number of visitors counted by the system and C_m represents the actual number of visitors based on manual observation. This equation measures the closeness of the automatic counting results to the ground-truth values across different test scenarios.

In addition to counting accuracy, two error-based metrics were used to provide a more detailed quantitative assessment of counting deviation, namely Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). The MAE was calculated to measure the average absolute difference between the number of visitors counted by the system and the ground-truth count, as shown in Equation (5). The MAPE was calculated to measure the average relative error between the system-generated count and the ground-truth count in percentage form, as shown in Equation (6). These metrics were selected because MAE represents the absolute magnitude of counting errors, whereas MAPE provides a scale-normalized error measure that facilitates comparison across videos with different visitor volumes [24], [25].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|C_{s,i} - C_{m,i}\right| \quad (5)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{C_{s,i} - C_{m,i}}{C_{m,i}}\right| \times 100\% \quad (6)$$

Where n denotes the number of test videos, $C_{s,i}$ represents the number of visitors counted by the system in the i-th video, and $C_{m,i}$ represents the corresponding ground-truth count obtained through manual observation. A lower MAE and MAPE indicate better counting performance. The evaluation was performed on videos from multiple tourist locations with different lighting conditions and levels of visitor density to assess the consistency of system performance in dynamic environmental settings.

2. Functional Testing

Functional testing was conducted using a black-box testing approach to verify whether all core system features operated according to the specified requirements from a user-facing perspective. The tested features included video upload, video playback control, automatic detection and tracking, visualization of bounding boxes, display of visitor count, display of video duration, and generation of processed video outputs. The expected and actual outputs were compared for each test scenario to determine the success status of each function.

3. System Performance Testing

System performance testing was conducted to assess the computational efficiency and real-time capability of the proposed system. The main performance indicators were average frame rate

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7759

and total processing time. The average frame rate was calculated using Equation (7).

$$FPS = \frac{N_f}{T_p} \quad (7)$$

Where N_f denotes the total number of processed frames and T_p represents the total processing time in seconds. A system was considered capable of real-time processing when the average FPS exceeded 25 frames per second, which was used as the minimum operational threshold in this study. The FPS and processing time results were used to evaluate whether the proposed system could support continuous video-based visitor monitoring in practical smart tourism environments.

RESULTS AND DISCUSSION

Visitor Counting Accuracy

The performance of the proposed YOLOv11-ByteTrack system was evaluated based on its ability to count visitors across nine test videos recorded at five tourist destinations in West Sumatra. The system-generated counts were compared with manually recorded ground-truth counts for each video. Table 4 presents the detailed counting results obtained using YOLOv3, YOLOv8, and YOLOv11 under the same evaluation setting.

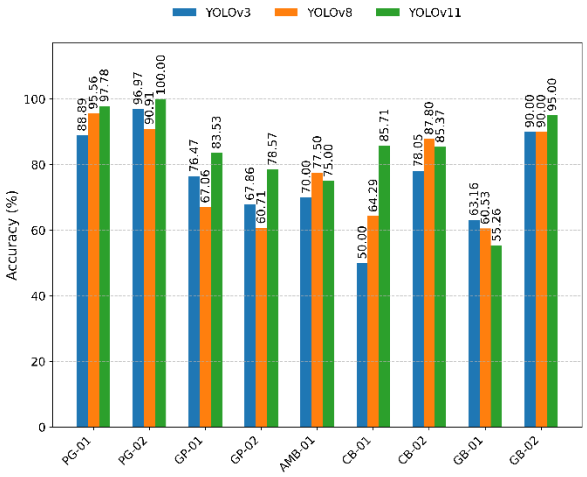
Table 4. Visitor Counting Accuracy Test Results

No	Name	Ground-truth	YOLOv3	YOLOv8	YOLOv11
1	PG-01	45	40	43	46
2	PG-02	33	34	36	33
3	GP-01	85	105	113	99
4	GP-02	56	74	78	68
5	AMB-01	40	52	49	50
6	CB-01	28	14	18	24
7	CB-02	41	32	46	47
8	GB-01	38	52	53	55
9	GB-02	20	18	18	19

Source: (Research Results, 2025)

As shown in Table 4, YOLOv11 produced the closest counts to the ground-truth values in six of the nine evaluated videos, namely PG-01, PG-02, GP-01, GP-02, CB-01, and GB-02. However, its performance was not uniformly superior across all locations. In AMB-01 and CB-02, YOLOv8 produced slightly closer counts, while in GB-01, YOLOv3 showed a smaller counting error than YOLOv11. This indicates that although YOLOv11 achieved the best overall performance, local scene characteristics such as camera viewpoint, background complexity, visitor movement near the counting line, and partial occlusion still influenced the final counting results.

Figure 4 presents the accuracy percentages of YOLOv3, YOLOv8, and YOLOv11 across the evaluated test videos. The figure provides a visual comparison of model performance under different outdoor tourism conditions, including differences in visitor density, illumination, and background complexity.



Source: (Research Results, 2025)

Figure 4. Accuracy Comparison of YOLOv3, YOLOv8, and YOLOv11 Across Tourist Destination Test Sites

Table 5 shows that YOLOv11 achieved the highest average counting accuracy of 84.02%, with the lowest MAE of 7.22 visitors per video and the lowest MAPE of 15.98%. In comparison, YOLOv3 achieved an average accuracy of 75.71%, with an MAE of 10.56 visitors per video and a MAPE of 24.29%, while YOLOv8 achieved an average accuracy of 77.15%, with an MAE of 10.67 visitors per video and a MAPE of 22.85%. These results indicate that YOLOv11-ByteTrack provided more accurate and stable counting estimates on average than the two comparison models under the evaluated outdoor tourism conditions.

Table 5. Summary of Counting Performance Metrics

No	Model	Average Accuracy	MAE (visitors/video)	MAPE
1	YOLOv3	75.71%	10.56	24.29%
2	YOLOv8	77.15%	10.67	22.85%
3	YOLOv11	84.02%	7.22	15.98%

Source: (Research Results, 2025)

Nevertheless, the detailed results also reveal that counting errors remained in several videos. The YOLOv11 model tended to produce overcounting in PG-01, GP-01, GP-02, AMB-01, CB-02, and GB-01, whereas undercounting occurred in CB-01 and GB-02. The largest discrepancy was observed in GB-01,

where YOLOv11 counted 55 visitors compared with the ground-truth value of 38. This suggests that overcounting may occur when the same visitor is detected or tracked inconsistently near the counting line, possibly due to ID switching, temporary occlusion, repeated movement around the line, or background complexity. Conversely, undercounting may occur when visitors are partially occluded, move too close to one another, or are not consistently detected across consecutive frames.

System Functionality Testing

System functionality testing was conducted using a black-box testing approach to verify whether the user-facing components of the system operated according to the expected specifications. The tested functions included video upload, playback control, video format description, visitor count display, video duration display, and processed video visualization. The results are summarized in Table 6.

Table 6. Functional Testing Results

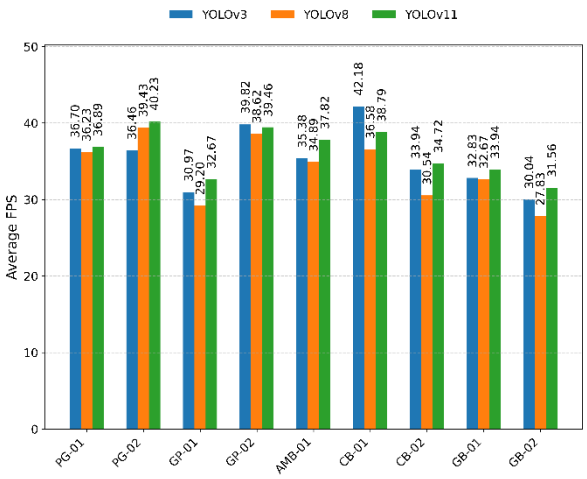
No	Test Scenario	Input / Function	Expected Output	Result
1	Start video button	Click "Start" button	Video starts playing	Success
2	Stop video button	Click "Stop" button	Video stops playing	Success
3	Video metadata display	Access description menu	Format information displayed	Success
4	Display visitor count	Run detection	Visitor count displayed	Success
5	Display video duration	Run detection	Duration displayed	Success
6	Display processed video	Run system	Output video displayed correctly	Success

Source: (Research Results, 2025)

As shown in Table 6, all tested functions operated successfully. These results indicate that the developed web-based system was functionally stable for video input handling, detection execution, tracking visualization, visitor count reporting, and processed video display. Although functional testing does not directly measure counting accuracy, it confirms that the system interface and processing workflow were able to support the experimental evaluation without operational failure.

System Performance Evaluation

System performance was evaluated based on the average frame rate achieved during video processing. Figure 5 presents the FPS comparison of YOLOv3, YOLOv8, and YOLOv11 across the evaluated test videos.



Source: (Research Results, 2025)

Figure 5. Average FPS Comparison of YOLOv3, YOLOv8, and YOLOv11 at All Test Sites

The system achieved an overall average processing speed of 36.23 FPS, with observed performance ranging from 27.83 FPS to 42.18 FPS. Since the observed FPS values exceeded the minimum real-time threshold of 25 FPS, the proposed system can be considered capable of real-time processing under the evaluated conditions. This finding is important because a visitor counting system for smart tourism management must not only produce accurate estimates but also process video streams with sufficiently low delay to support operational decision-making.

The FPS results also indicate that YOLOv11 maintained competitive processing speed while producing the highest average counting accuracy among the evaluated models. This suggests that the integration of YOLOv11 and ByteTrack provides a practical trade-off between accuracy and computational efficiency. However, FPS alone does not guarantee counting reliability. In visually complex scenes, the system may still produce counting errors despite maintaining real-time processing speed. Therefore, accuracy and processing speed should be interpreted jointly when assessing the feasibility of the proposed system for real-world tourism monitoring.

Interpretation of Findings

The experimental results show that the YOLOv11-ByteTrack system is feasible for real-time visitor counting in outdoor tourism environments, but its performance remains dependent on scene characteristics. The average accuracy of 84.02%, combined with an average processing speed of 36.23 FPS, indicates that the system provides a reasonable balance between counting accuracy and real-time computational efficiency. This balance is

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7759

particularly relevant for smart tourism applications where destination managers require timely information on visitor flow, capacity conditions, and crowd movement.

The strongest performance was observed in PG-01 and PG-02, where YOLOv11 achieved 97.78% and 100% accuracy, respectively. These results suggest that the system performed well when visitors were sufficiently visible and their movement across the counting line was relatively clear. Under such conditions, the combination of YOLOv11 detection and ByteTrack identity association was able to maintain stable object trajectories and reduce repeated counting.

However, the results also show that YOLOv11 did not produce the best result in all test videos. In AMB-01 and CB-02, YOLOv8 produced slightly closer counts to the ground truth, while in GB-01, YOLOv3 showed a lower counting error than YOLOv11. This finding indicates that newer object detection models do not automatically guarantee superior counting performance in every scene. In visitor counting tasks, the final count is affected not only by detection capability but also by tracking stability, object trajectory continuity, counting-line placement, and visitor behavior near the counting area.

The most critical error pattern observed in YOLOv11 was overcounting. For example, in GP-01, YOLOv11 counted 99 visitors compared with the ground-truth value of 85, while in GB-01 it counted 55 visitors compared with the ground-truth value of 38. These discrepancies suggest that some visitors may have been counted more than once due to tracking instability, ID switching, temporary loss of object identity, or repeated movement near the virtual line. Although ByteTrack is designed to preserve object identity across frames, tracking errors may still occur when visitors overlap, become partially occluded, or move through visually complex backgrounds.

Undercounting was also observed, although less frequently. In CB-01, YOLOv11 counted 24 visitors compared with the ground-truth value of 28, while in GB-02 it counted 19 visitors compared with the ground-truth value of 20. These errors may be associated with missed detections, partial occlusion, small object size, or visitor movement outside the optimal field of view. This indicates that both detection failure and tracking instability can affect counting performance, depending on the visual characteristics of each scene.

Lighting and background complexity also influenced the system performance. Outdoor tourism environments are more difficult than controlled indoor environments because illumination may change due to weather, time of day, shadows, and camera orientation. In addition, natural and coastal tourism backgrounds often contain visually complex elements that may reduce detection confidence or interfere with consistent tracking. Therefore, the achieved accuracy should be interpreted as a practical baseline for a pre-trained YOLOv11-ByteTrack system under real outdoor tourism conditions, rather than as the maximum possible performance after domain-specific optimization.

Overall, the results demonstrate that YOLOv11-ByteTrack is suitable for real-time visitor counting in the evaluated outdoor tourism scenarios, particularly when visitor trajectories are clear and occlusion is limited. However, the system remains sensitive to overlapping visitors, complex backgrounds, camera placement, and repeated movement near the counting line. These findings indicate that future improvements should focus not only on increasing detection accuracy but also on improving tracking robustness, optimizing counting-line placement, and reducing ID switching in crowded or visually complex scenes.

Comparison with Previous Studies

The findings of this study are generally consistent with previous research showing that deep learning-based object detection models can support people counting, crowd monitoring, and visitor analytics from video or image data [16], [20], [21], [22]. YOLO-based approaches have been widely used because they offer a practical balance between detection accuracy and processing speed, making them suitable for real-time monitoring tasks [21], [22], [26], [27]. The present study supports this observation by showing that YOLOv11-ByteTrack achieved an average accuracy of 84.02% while maintaining an average processing speed of 36.23 FPS.

However, the present findings should not be interpreted as directly equivalent to results reported in previous studies because the evaluation context is different. Several prior studies on people counting, crowd detection, or occupancy monitoring were conducted in public spaces, indoor environments, or relatively structured settings where lighting conditions, camera viewpoints, and movement patterns were more stable [20], [22], [23]. By contrast, this study evaluated the system in open-air tourist destinations, where illumination changes, background complexity, visitor movement, and occlusion are more difficult to control. This difference explains why the achieved accuracy is meaningful as field-based evidence, even though counting errors remained in several scenes.

Compared with previous tourism-related computer vision studies, this research extends the evaluation context toward real-time visitor counting in outdoor tourist destinations. Camera-based visitor monitoring has been explored in natural and recreational environments [16], while recent work has also demonstrated the potential of YOLO-based tourist crowd monitoring for overtourism mitigation [21]. The present study differs by evaluating a YOLOv11-ByteTrack detection and tracking pipeline across multiple open-air tourist destinations in West Sumatra, using videos with different lighting conditions and visitor densities. Therefore, the contribution of this study lies not only in the achieved accuracy but also in demonstrating how a real-time detection-tracking system performs under less controlled tourism field conditions.

The limitations observed in this study also align with findings from previous multi-object tracking and crowd-counting research. ByteTrack improves tracking continuity by associating detection boxes across frames [28], but it cannot fully eliminate errors caused by occlusion, overlapping objects, and identity switches. Similarly, crowd-counting studies have reported that dense scenes, visual obstruction, and scale variation remain major sources of counting error [29]. The overcounting observed in GP-01 and GB-01 supports this limitation, indicating that tracking-assisted counting can still be affected when object identities are fragmented or when visitors move repeatedly near the counting line.

Another important distinction is that the YOLOv11 model in this study was used without fine-tuning on local tourism data. This differs from studies that optimize models using domain-specific datasets. As a result, the reported performance should be understood as a baseline for a readily deployable pre-trained model rather than an optimized upper-bound result. This has practical value because tourism managers may not always have access to large annotated local datasets, but it also suggests that future work may improve accuracy through local dataset expansion, fine-tuning, camera placement optimization, and more robust re-identification mechanisms.

Overall, compared with previous studies, the present research provides additional evidence that YOLO-based detection combined with multi-object tracking can support real-time visitor counting in tourism environments. At the same time, the findings show that outdoor tourist destinations introduce specific challenges that are less dominant in controlled settings, particularly illumination variation, occlusion, repeated movement near counting lines, and scene-specific background complexity. These issues should be addressed in future research to improve the reliability of AI-based visitor monitoring systems for smart tourism management.

Theoretical and Practical Implications

Theoretically, this study contributes to smart tourism research by demonstrating that a detection-tracking-counting pipeline can be applied to visitor analytics in uncontrolled outdoor tourism environments. Rather than evaluating computer vision only in controlled or generic surveillance settings, the present study shows how YOLOv11 and ByteTrack perform when exposed to real tourism conditions involving variable lighting, complex backgrounds, and non-uniform visitor movement. This strengthens the empirical basis for integrating AI-based computer vision into smart destination monitoring frameworks.

The findings also indicate that the combination of object detection and multi-object tracking is important for visitor counting tasks. Object detection alone is insufficient when visitors move across frames because the same person may be detected repeatedly. By integrating ByteTrack, the system can maintain object identity continuity and reduce duplicate counting. Nevertheless, the remaining errors show that tracking continuity is still vulnerable to occlusion, overlapping visitors, and ID switching. This provides a methodological insight that visitor counting accuracy depends on both detection quality and tracking stability.

From a practical perspective, the system can support destination managers in monitoring visitor flow, assessing capacity conditions, and identifying changes in crowd movement. The average processing speed of 36.23 FPS indicates that the system has the potential to operate in near real-time, which is important for operational decisions such as adjusting staff allocation, managing entry flow, and responding to crowd accumulation. Because the system was tested on a mid-range computing platform, it also shows potential for cost-effective implementation by regional tourism agencies.

However, practical deployment should consider camera placement, counting-line position, and site-specific movement patterns. The results indicate that errors may increase when visitors overlap, move repeatedly near the counting line, or pass through visually complex backgrounds. Therefore, implementation in real tourist destinations should be accompanied by site-specific calibration, careful camera positioning, and periodic

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7759

validation against manual counts to ensure operational reliability.

Limitations and Future Work

Despite the encouraging results, several limitations should be considered when interpreting the findings. First, the evaluation was conducted using nine videos recorded at five tourist destinations in West Sumatra. Although the dataset includes variation in lighting, visitor density, and environmental background, it remains limited in scale and geographic coverage. Therefore, the findings should be interpreted as evidence of feasibility in the evaluated settings rather than as a generalized representation of all outdoor tourism environments.

Second, the ground-truth counts were obtained through manual observation by a single observer. This approach provided a practical reference for evaluation but may introduce subjectivity, especially in scenes involving partial occlusion, overlapping visitors, or ambiguous visual boundaries. Future studies should involve multiple annotators and inter-observer reliability checks to improve the validity of ground-truth data.

Third, the YOLOv11 model was used in its off-the-shelf configuration without fine-tuning on local tourism data. This choice was useful for evaluating baseline deployability, but it may have limited detection performance in scenes with domain-specific visual characteristics. Future research should investigate whether fine-tuning with locally annotated tourism datasets can reduce counting errors, especially under dense crowd conditions and complex outdoor backgrounds.

Fourth, although ByteTrack helped maintain identity consistency across frames, counting errors remained when visitors overlapped, moved repeatedly near the counting line, or experienced temporary loss of identity. Future work should explore more robust tracking strategies, including re-identification-based tracking, transformer-based tracking, multi-camera configurations, and adaptive counting-line placement. Integration with destination management dashboards may also be explored to support practical smart tourism applications such as real-time visitor flow visualization, capacity warning systems, and evidence-based infrastructure planning.

CONCLUSION

The integration of YOLOv11 and ByteTrack produced a real-time video-based visitor counting system that achieved an average counting accuracy of 84.02%, MAE of 7.22 visitors per video, MAPE of 15.98%, and an average processing speed of 36.23 FPS on nine Full HD videos from five tourist destinations in West Sumatra. These findings show that YOLOv11–ByteTrack provided the best overall average performance compared with YOLOv3 and YOLOv8 in the evaluated scenarios, while maintaining real-time processing capability under varying lighting conditions and visitor densities. The results indicate that the proposed system can support smart tourism management by providing real-time visitor flow information for capacity assessment, crowd monitoring, safety management, staff allocation, and data-driven infrastructure planning, particularly for regional destinations requiring cost-effective monitoring solutions. Nevertheless, counting errors still occurred in visually complex scenes, especially when visitors overlapped, became partially occluded, or moved repeatedly near the counting line, which may lead to overcounting or undercounting due to detection errors, tracking instability, or identity switching. Future research should improve system reliability under denser and more complex crowd conditions by expanding the dataset, involving multiple annotators for stronger ground-truth validation, applying local fine-tuning, improving tracking through re-identification or spatio-temporal association, exploring multi-camera configurations, and integrating the system into destination management platforms such as real-time dashboards and capacity alert systems.

ACKNOWLEDGMENT

The authors would like to thank Lembaga Penelitian dan Pengabdian Masyarakat Universitas Negeri Padang for funding this work with a contract number: 2032/UN35.15/LT/2025.

REFERENCE

[1] I. Sustacha, J. F. Baños-Pino, and E. del Valle, "The role of technology in enhancing the tourism experience in smart destinations: A meta-analysis," *Journal of Destination Marketing & Management*, vol. 30, p. 100817, 2023, doi: 10.1016/j.jdmm.2023.100817.

[2] S. Bingöl and Y. Yang, "Integrating smart technologies and artificial intelligence to build smart tourism destination ecosystems: A model for smart destination management," *Tourism Management Perspectives*, vol. 58, p. 101380, 2025, doi: 10.1016/j.tmp.2025.101380.

[3] M. A. Celdrán-Bernabéu, J.-N. Mazón, D. Giner-Sánchez, J. Morales-García, and M. P.

Peñarrubia-Zaragoza, "Smart tourism destinations as open data providers: Barriers and opportunities," *Journal of Destination Marketing & Management*, vol. 40, p. 101065, 2026, doi: 10.1016/j.jdmm.2025.101065.

[4] C. N. Novera, Z. Ahmed, R. Kushol, P. Wanke, and Md. A. K. Azad, "Internet of Things (IoT) in smart tourism: A literature review," *Spanish Journal of Marketing - ESIC*, vol. 26, no. 3, pp. 325–344, 2022, doi: 10.1108/SJME-03-2022-0035.

[5] J.-W. Bi, C. Li, H. Xu, and H. Li, "Forecasting daily tourism demand for tourist attractions with big data: An ensemble deep learning method," *Journal of Travel Research*, vol. 61, no. 8, pp. 1719–1737, 2022, doi: 10.1177/00472875211040569.

[6] M. Mariani and R. Baggio, "Big data and analytics in hospitality and tourism: A systematic literature review," *International Journal of Contemporary Hospitality Management*, vol. 34, no. 1, pp. 231–278, 2022, doi: 10.1108/IJCHM-03-2021-0301.

[7] Y. Cai, G. Li, L. Wen, and C. Liu, "Intellectual landscape and emerging trends of big data research in hospitality and tourism: A scientometric analysis," *International Journal of Hospitality Management*, vol. 117, p. 103633, 2024, doi: 10.1016/j.ijhm.2023.103633.

[8] S. Park, "Big data in smart tourism: A perspective article," *Journal of Smart Tourism*, vol. 1, no. 3, pp. 3–5, 2021, doi: 10.52255/smarttourism.2021.1.3.2.

[9] D. P. Sakas, D. P. Reklitis, M. C. Terzi, and C. Vassilakis, "Multichannel digital marketing optimizations through big data analytics in the tourism and hospitality industry," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 17, no. 4, pp. 1383–1408, 2022, doi: 10.3390/jtaer17040070.

[10] J. B. Read, M. Daniels, and L. Harmon, "Implementing technology-based visitor counts in parks: A methodological overview," *Journal of Park and Recreation Administration*, vol. 39, no. 1, pp. 85–103, 2021, doi: 10.18666/JPRA-2020-10502.

[11] M. J. Daniels, H.-L. Liu, and S. L. Powers, "Infrared visitor counts: Data validation and algorithm development," *Current Issues in Tourism*, vol. 28, no. 14, pp. 2215–2219, 2025, doi: 10.1080/13683500.2024.2364758.

[12] D. Schmücker *et al.*, "The INPReS intervention escalation framework for avoiding overcrowding in tourism destinations," *Tourism and Hospitality*, vol. 4, no. 2, pp. 282–292, 2023, doi: 10.3390/tourhosp4020017.

[13] J. Reif, D. Schmücker, L. Naschert, and E. Horster, "Visitor management in tourism destinations: Current challenges in measuring and managing visitors' spatio-temporal behaviour," in *Tourism Destination Development: A Geographic Perspective on Destination Management and Tourist Demand*, M. Pillmayer, M. Karl, and M. Hansen, Eds. Berlin, Germany: De Gruyter, 2024, pp. 81–104, doi: 10.1515/9783110794090-005.

[14] S. Zhang and A. Chen, "Do different queue formations influence the overestimation of tourism carrying capacity?," *Sustainability*, vol. 16, no. 24, p. 11047, 2024, doi: 10.3390/su162411047.

[15] G. Lupp *et al.*, "Visitor counting and monitoring in forests using camera traps: A case study from bavaria (Southern Germany)," *Land*, vol. 10, no. 7, p. 736, 2021, doi: 10.3390/land10070736.

[16] J. Staab, E. Udas, M. Mayer, H. Taubenböck, and H. Job, "Comparing established visitor monitoring approaches with triggered trail camera images and machine learning based computer vision," *Journal of Outdoor Recreation and Tourism*, vol. 35, p. 100387, 2021, doi: 10.1016/j.jort.2021.100387.

[17] A. Panigrahy and A. Verma, "Tourist experiences: A systematic literature review of computer vision technologies in smart destination visits," *Journal of Tourism Futures*, vol. 11, no. 2, pp. 187–202, 2025, doi: 10.1108/JTF-04-2024-0073.

[18] X. Wang, "Construction of smart tourism system integrating tourist needs and scene characteristics," *Systems and Soft Computing*, vol. 6, p. 200168, 2024, doi: 10.1016/j.sasc.2024.200168.

[19] D. Nurseitov, K. Bostanbekov, N. Toiganbayeva, A. Zhalgas, D. Yedilkhan, and B. Amirgaliyev, "Vision-based people counting and tracking for urban environments," *Journal of Imaging*, vol. 12, no. 1, p. 27, 2026, doi: 10.3390/jimaging12010027.

[20] N. Krishnachaithanya *et al.*, "People counting in public spaces using deep learning-based object detection and tracking techniques," in *2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, Greater Noida, India, 2023, pp. 784–788, doi: 10.1109/CISES58720.2023.10183503.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7759

[21] K. Wijayanti, G. A. Mutiara, B. Suryawardani, E. Ervina, and G. P. Kusuma, "Non-intrusive real-time tourist crowd monitoring for overtourism mitigation using YOLOv8-based head detection and tracking," *Journal of Robotics and Control (JRC)*, vol. 6, no. 4, pp. 1985–2004, 2025, doi: 10.18196/jrc.v6i4.26396.

[22] M. Ş. Gündüz and G. Işık, "A new YOLO-based method for real-time crowd detection from video and performance analysis of YOLO models," *Journal of Real-Time Image Processing*, vol. 20, no. 1, p. 5, 2023, doi: 10.1007/s11554-023-01276-w.

[23] W. Zhang, J. Calautit, P. W. Tien, Y. Wu, and S. Wei, "Deep learning models for vision-based occupancy detection in high occupancy buildings," *Journal of Building Engineering*, vol. 98, p. 111355, 2024, doi: 10.1016/j.jobe.2024.111355.

[24] A. Jierula, S. Wang, T.-M. Oh, and P. Wang, "Study on accuracy metrics for evaluating the predictions of damage locations in deep piles using artificial neural networks with acoustic emission data," *Applied Sciences*, vol. 11, no. 5, p. 2314, 2021, doi: 10.3390/app11052314.

[25] H. Khoshvaght, R. R. Permala, A. Razmjou, and M. Khiadani, "A critical review on selecting performance evaluation metrics for supervised machine learning models in wastewater quality prediction," *Journal of Environmental Chemical Engineering*, vol. 13, no. 6, p. 119675, 2025, doi: 10.1016/j.jece.2025.119675.

[26] R. Khanam and M. Hussain, "YOLOv11: An overview of the key architectural enhancements," arXiv preprint, arXiv:2410.17725, 2024, doi: 10.48550/arXiv.2410.17725.

[27] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, 2023, pp. 7464–7475, doi: 10.1109/CVPR52729.2023.00721.

[28] Y. Zhang *et al.*, "ByteTrack: Multi-object tracking by associating every detection box," in *Computer Vision – ECCV 2022*, Cham: Springer Nature Switzerland, 2022, pp. 1–21, doi: 10.1007/978-3-031-20047-2_1.

[29] L. Deng, Q. Zhou, S. Wang, J. M. Górriz, and Y. Zhang, "Deep learning in crowd counting: A survey," *CAAI Transactions on Intelligence Technology*, vol. 9, no. 5, pp. 1043–1077, 2024, doi: 10.1049/cit2.12241.