TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 23/23 pp2304–2321
DOI: 10.26599/TST.2024.9010003
Volume 31, Number 4, August 2026

Research Article

MU-Net-optLSTM: Two-Stream Spatial–Temporal Feature Extraction and Classification Architecture for Automatic Monitoring of Crowded Art Museums

Mukun Wang, Rongju Yao, and Khosro Rezaee*

Abstract: Networked cameras that continuously capture video data have generated a high demand for hybrid edge-to-cloud servers that can process live videos in real time. The environment of art museums is rarely studied, but visual analysis is an important factor in categorizing and distinguishing individuals and crowds through smart surveillance systems. This paper demonstrates how video surveillance data from art museums can be analyzed to identify abnormal behavior using an innovative deep learning framework. To enhance the extracted features, a spatial feature extraction method based on the U-Net architecture is applied, along with the encoder component of the proposed approach, MobileNetV2. Additionally, we propose an improved Long-Short-Term Memory (LSTM) algorithm for extracting temporal features. Optical flow enhances surveillance in art museums by tracking individuals and crowds. Our approach yields an average accuracy of 97.67±1.23% when applied to a collection of video datasets. Using U-Net, MobileNetV2, and optimized LSTM algorithms, the model recognizes patterns in video data, such as crowd motion in museums. Consequently, this methodology generates reliable results as well as being computationally efficient. Compared to the state-of-the-art, the proposed method is more comprehensive and generalizable for analyzing atypical museum visitor behavior.

Key words: art museums; automatic monitoring; deep learning; long short-term memory; U-Net architecture; MobileNetV2; optical flow

1 Introduction

Efficient Video surveillance technology is employed by security professionals to evaluate collective behavior in public spaces[1, 2]. In recent years, the significance of employing many smart cameras in an edge-fog-cloud infrastructure for continuous surveillance of individuals and crowds has increased[3, 4]. Data-driven crowd simulations[4, 5] and advanced driving assistance systems[6, 7] heavily depend on multi-person and multi-camera tracking techniques. Among the significant places that require surveillance and video analysis are art museums[8]. Insufficient research has been conducted to assess the effectiveness of automated video surveillance and crowd analysis in art museum environment However, art museums and other organizations responsible for safeguarding precious artifacts are legally obligated to implement all essential measures to protect data confidentiality[9]. Maintaining records of surveillance actions inside monitored zones

• Mukun Wang is with Department of Design, Graduate School, Dongseo University, Busan 47011, Republic of Korea, and also with Tianshui Normal University, Tianshui 741000, China. E-mail: miachiyu2022@gmail.com.
• Rongju Yao is with Shandong Provincial University Laboratory for Protected Horticulture, Weifang University of Science and Technology, Weifang 262700, China. E-mail: yaorongju8@wfust.edu.cn.
• Khosro Rezaee is with Department of Biomedical Engineering, Meybod University, Yazd 89195, Iran. E-mail: kh.rezaee@meybod.ac.ir.
* To whom correspondence should be addressed.
Manuscript received: 2023-11-24; revised: 2023-12-18; accepted: 2023-12-29

is difficult, and recording and preservation for future use is impractical. Nevertheless, the exploration of automated learning techniques for monitoring visitors and classifying them according to their emotional responses has not been thoroughly investigated. This is to protect invaluable antiquities[10].

Identifying and comprehending atypical behavior among large populations, such as art museums, is a challenge[11]. Most tracking frameworks undergo evaluation and refinement using offline video data. A dynamically lit environment facilitates tracking visitors and their behaviors due to fluctuating light levels. This aspect is frequently overlooked in the literature pertaining to multi-object tracking (MOT), as seen in Refs. [12, 13]. Various Camera Tracking (MCT) methodologies[14], employ clustering or camera model generation techniques to merge tracklets obtained from multiple cameras. MOT systems improve tracking accuracy when deep features are extracted from each motion frame[15]. Recent findings have indicated that the establishment of a comprehensive categorization system for each visitor's physical characteristics can facilitate the identification of correlations between documented visits and incoming visitors[16, 17]. Deep learning methodology with visitor monitoring enhances multi-camera tracking in multiplayer mode[18]. Art museums have the potential to enhance cost-efficiency and optimize surveillance of expansive exhibition spaces through visitor tracking and identification systems, without increasing the number of cameras employed.

Online methods are employed to effectively classify visitors and establish connections between their present and past visits. As a result, art museums are less likely to experience theft, overcrowding, or abnormal behavior[19]. The aforementioned served as inspiration for our flexible deep learning technology. A comparable method has been implemented in our competitive strategy to track and analyze visitors to the art museums based on the expensive items in the museum as high-risk and crowded environments. Consequently, a novel approach has been developed to enhance multi-camera tracking flexibility and reliability, while reducing superfluous data. The adoption of visitor monitoring and identification systems by art museums has the potential to yield two significant outcomes: cost savings and enhanced security measures within large exhibition rooms.

This article is primarily concerned with tracking people's visits to crowded places and recognizing unusual behavior in art museums. By processing received videos and training on a large number of frames from normal and abnormal populations, this issue is addressed. Next, the proposed method will be used as a ready-made pattern, applicable to online and real-time applications. The act of recognizing and tracking human or congestion here refers to crowds whose irregular movements and actions can lead to crimes and damage to art museum objects. With Internet data, we made the model most congested with normal visitor conditions and abnormal demographic conditions such as obstacles, suspicious objects, panic, fighting, and congestion.

As demonstrated in this study, spatial information can be extracted using a methodology similar to the U-Net architecture. In our proposed model, MobileNetV2 represents the encoder component. In addition, we introduce an improved version of the Long-Short-Term Memory (LSTM) methodology for extracting temporal characteristics. To enhance surveillance and improve classification, optical flow is employed at the end of the process to track people and crowds within art museums. The average accuracy of 97.67 ± 1.23% was achieved when this methodology was applied to authentic surveillance footage. The results generated by this approach are reliable and demonstrate computational efficiency. Drawing upon this foundational principle, we also employed edge computing[20, 21] to facilitate museum video processing, thereby reducing the distance between computational resources and data sources. The relocation of computing resources to the edge can yield several benefits, including enhanced application performance, decreased data transportation expenses, and the ability to manage public cloud expenditures.

- To our knowledge, no automated video-based surveillance system has been proposed for art museums that can analyze online video frames for overcrowding and security processes in recognizing abnormal visitor behavior.
- The proposed method has the potential to identify instances of crowd overcrowding within densely populated art museums through the monitoring and analysis of visitor movements. The innovative design facilitates rapid recognition while maintaining high precision levels. Furthermore, we utilized MobileNetV2's encoder and the U-Net architecture to extract spatial features. Moreover, an enhanced

iteration of the LSTM method is proposed for acquiring temporal characteristics. Consequently, the method depends on a rapid categorization model for making real-time judgments in scenarios characterized by a significant concentration of individuals.

- The proposed approach has the potential to be utilized in art galleries that see high levels of foot traffic. The utilization of deep learning in congestion management strategies has the potential to address irregular human and crowd behavior.
- The enhanced model significantly enhances the foundational decision-making framework. Two markers of system strength include overcrowding and atypical human behavior.

The initial idea of the research originated from the need to address security concerns in art museums and explore the potential of utilizing video surveillance data. The authors were motivated to investigate this topic to develop a novel deep learning framework that could analyze the data and identify abnormal behavior. By combining spatial feature extraction using the U-Net architecture and the MobileNetV2 encoder, along with an improved LSTM algorithm for extracting temporal features, the authors aimed to enhance surveillance in art museums by effectively tracking individuals and crowds using optical flow. Art museums are characterized by two factors: overcrowding and safeguarding valuable possessions. As a result, video processing techniques for crowd recognition have the potential to reduce continuous monitoring, presenting a significant advantage. Congestion monitoring provides several benefits in various aspects.

The subsequent sections of this study comprise the main body of the paper. Section 2 describes the literature review. In Section 3, a more comprehensive exploration of the proposed paradigm is outlined. Additionally, datasets, results, and commentary are presented and discussed in Section 4. In section 5, we present the discussion part of the paper. The report concludes by offering recommendations for further research based on the analysis findings.

2 Related Work

With tracking technology, visitors' trajectories within an art museum can be mapped. The authors proposed using Light Detection and Ranging (LiDAR) technology to recognize and track humans, as well as determine their direction and trajectory[10]. In any unobstructed area, this phenomenon may be observed. When individuals enter an art museum, technology records their location. Located in Kurashiki, Okayama Prefecture, Japan, the Ohara Museum of Modern Art strives to achieve this objective. Many advocate the integration of visitor tracking technologies into mobile devices[22, 23]. To identify and assess user movement patterns, previous studies employed sophisticated and expensive monitoring technologies[24–26]. Lanir et al.[24] introduced a visual aid designed for art museum staff. Using the program, users' locations and actions can be monitored and recorded, allowing frequently visited areas to be identified. The SeeForMe approach[25] is an alternative. A computer vision framework can be used in real-time on mobile devices to recognize and classify a variety of objects and artistic creations. An audio tour incorporates a video camera to facilitate the identification and labelling of artworks. An audio tour accompanied by a visual system was evaluated at the Bargello Museum in Florence as part of a pilot initiative. To recognize and classify objects, a Convolutional Neural Network (CNN)[26] is augmented with NVIDIA mobile Graphics Processing Units (GPUs). During the tour, a visitor can engage in a discussion that temporarily halts the tour. As a result of the speech recognition module, which takes several parameters into account, such as companionship, this interruption may be identified. A training dataset of 300 individuals and their corresponding photographic images was used in the study. Optimal accuracy and recall rates are observed when distances are limited to 5 meters or less. With only 22 instances of error, the researchers achieved near-perfect recognition for the majority of works after modifying the algorithm. Participants on the System Usability Scale expressed concerns about the invasive nature of the guide throughout their visit. According to the System Usability Scale (SUS), the system is highly usable. Also, the camera must be kept in a shirt pocket positioned at chest level, which may seem cumbersome.

The study presented in Ref. [27] uses an RGB-D camera and a computer vision technique based on Kinect technology. To identify influential opinion leaders and analyze group behavior, the National Museum of Emerging Science and Innovation in Japan meticulously observes its visitors. Based on Kinect and RGB-D cameras,[27] presents a computer vision methodology in their study. Scholars have been

monitoring and evaluating the dynamics of groups of individuals visiting art museums to better understand how these collectives are formed and operated. Nonverbal cues and leader spatial orientation are considered. This study explores whether robotic museum docents could replace human tour guides as a primary objective. Across the grounds of the National Museum of Emerging Science and Innovation in Tokyo, Japan, four Kinect V1 sensors have been strategically positioned. To capture interactions between docents and tour groups, video recordings were made over a two-month period. In order to determine the direction of motion, disruptions in the bounding boxes of consecutive photographs can be analyzed. During the experiment, the guide and participants were required to manually annotate their movements and compare them to the algorithm's predictions. As a result of the close proximity of individuals to the camera, occlusion and imprecise measurements are the primary challenges associated with this approach. Using the exponential motion approach, it may also be possible to increase the accuracy of bounding box classification. This is currently estimated at 70–75%.

The authors describe an Internet of Things (IoT)-based system that monitors and evaluates visitor behavior within the Galleria Borghese in Rome, Italy[28]. An analysis of a case study conducted at the CoBrA Museum of Modern Art in Amstelveen, the Netherlands, is presented in Ref. [29]. Numerous studies have investigated prevalent methods for indoor localization[30–33].

Established approaches to user timing and tracking are based on the progressive development of RGB video models and processes. In motion analysis, motion capture, and other virtual reality (VR) endeavors, the aforementioned techniques are extensively employed[34]. Analyzing visual data obtained from RGB cameras can facilitate the precise location of the visitor in this context. This objective can be achieved by employing two distinct capture methods. In the first approach, visitor information is obtained from a database, whereas the second technique relies on art-specific datasets. Strategic camera placement determines their effectiveness.

According to the study[35], previous research efforts employed computer vision techniques and content-based photo retrieval methods. Obtaining a visitor ID involves first utilizing object recognition technology, then analyzing and representing the acquired data using deep learning methodology. Many cameras strategically placed throughout the exhibition space support this strategy. Distances between distributions are used to measure color similarity in the visual classification of individuals. To enhance efficiency and save time, the current system requires optimization.

It has been shown in study[36] that the Kinect sensor can be used to track objects. They analyze data pertaining to the head and face orientation to determine eye contact. Using face direction data and object identification, the authors present a methodology for gaze and object prediction. By monitoring head and eye movements simultaneously with a Kinect sensor, researchers may be able to estimate individuals' focal points of attention.

A methodology similar to ours is shown in Ref. [37]. It is intended to accumulate a lot of data related to user engagement. This includes visitor trajectories, entrance counts, pedestrian flow, and the duration of artwork observation. Using infrared cameras and re-identification algorithms, the researchers integrated security cameras with infrared technology. For human re-identification in video, our proposed methodology significantly differs from conventional methods during the preprocessing phase. Prior to its presentation to the user, the aforementioned procedure is performed directly on the badge itself through our system. In comparison with other tracking methods, the approach described allows real-time visitor tracking. This differentiation enables the development of innovative tools that benefit both visitors (via recommender systems) and art museum curators and staff (via visitor flow analysis[38]).

In Ref. [39] a study focused on deep learning to determine damage to ancient masonry buildings. The study used mobile deep learning to identify areas of damage in historic buildings based on images captured from a mobile platform. This method can quickly and efficiently recognize damage to historic buildings. The disadvantage is that it is limited in accuracy and may not recognize certain types of damage, such as structural degradation. Additionally, this method may not be suitable for all historic buildings, as it requires certain technological infrastructure.

The study[40] proposes to create a system that uses data from museums and exhibitions to recommend items to visitors. This would help people find items that interest them and make museums more efficient.

The advantage of the system is that it could help people find items that interest them, making the museums more efficient. However, the system could also lead to visitors feeling overwhelmed by too many choices or not knowing which items to spend their time on. Additionally, the system could be vulnerable to data breaches if the data were not properly protected.

In Ref. [41], deep learning is applied to predict museum visitors' artwork preferences. Deep learning algorithms classify artwork using museum visitors' surveys data. According to the study, the algorithm predicted visitors' preferences accurately. This approach provides insights into visitor behavior and preferences. This can help museum curators better understand their visitors' needs and tailor their exhibitions to them. One disadvantage is that it can be time-consuming and costly to collect the data needed to train the algorithm, as well as to implement the algorithm in the museum's system.

Despite this, the potential of this approach in predicting visitors' preferences makes it a valuable tool for museum curators. Ultimately, the implementation of deep learning for smart video security and automated monitoring of crowds and humans in museums can provide valuable insight into visitor behavior. This can help create a more engaging museum experience.

3 Methodology

The proposed model aims to match the performance of current solutions in the realm of crowd behavior recognition and tracking the movement of large groups of individuals. This is intended to prevent potential harm to art museum displays. In an effort to streamline the computational process for use in regulatory devices, we conducted a comparative analysis between our suggested methodology and the current state-of-the-art. The proposed method consists of three sequential steps: Firstly, the utilization of time-distributed U-Nets or spatial feature extraction techniques is considered. Next step involves temporal information extraction, followed by categorization. In addition, as a result of these three steps, optical flow is able to track people's movements in the art museum. This study aims to design a convolutional classifier that exhibits performance similar to that of established state-of-the-art models, while also demonstrating computational efficiency. This technique is considered appropriate for real-time operations on devices with limited bandwidth because of its decreased memory needs.

3.1 MobileNetV2 and U-Net architectures

The model includes a brief segment of security camera video lasting one second, with 30 individual frames. Upon receiving a video frame, the model initiates U-Net for spatial feature extraction. Figure 1 illustrates the utilization of MobileNetV2 as an encoder inside the U-Net architecture to extract static single-frame spatial attributes from a sequential temporal distribution.

The U-Net architecture uses a combination of an encoder and a decoder, with MobileNetV2 serving as the encoder. The encoder is used to extract static

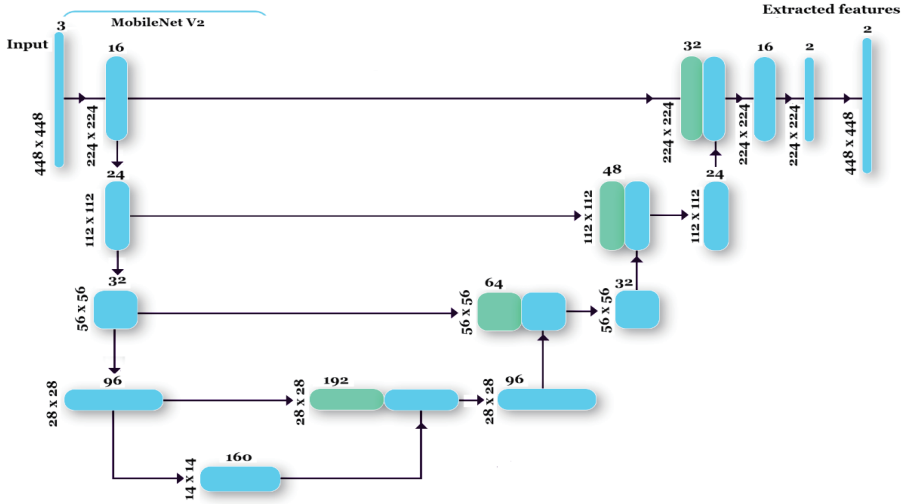

Fig. 1 Visual representations of MobileNet V2 and U-Net networks. Feature maps for both the decoder and encoder, which are MobileNet V2 maps, are depicted in the green figure.

single-frame spatial attributes from the video sequence, which are then fed into the decoder to produce the final output. The encoder is used to transform the input video sequence into a set of features, while the decoder is used to reconstruct the output video sequence from these extracted features.

The MobileNetV2 architecture is used in the encoder because it is an efficient and high-performing neural network that can be used to produce accurate features from the input video sequence.

3.2 Optimized LSTM

The utilization of the LSTM technique[42] enhances the effectiveness of the collaborative suggestion strategy by facilitating accurate rating prediction. In order to achieve accurate prediction using multi-layer LSTM models, it is essential to provide a comprehensive set of input features. The $p_{i,j}$= fΘ(Feature) and fΘ layers in LSTM networks serve as prediction layers and generate predictions. The LSTM network is represented by the symbol f, whereas the projected rating for output *j*, given user *i*, is indicated by p_{ij}.

$$\text{Feature}_{u_i}^{e_j} = \begin{pmatrix} FE(u_i), FE(e_j), TE(u_i), IE(u_i), IE(e_j), \\ NF(u_i), NF(e_j), NT(u_i) \end{pmatrix} \tag{1}$$

In this context, the variable m refers to the total count of concatenated attributes, namely 8. Additionally, the symbol E_d represents the embedding representation of the *d*-th feature. The input concatenated embedding of the LSTM is expressed by Feature $^{(0)}$ = [E_1, E_2, E_3, …, Em].

LSTM networks have emerged as the prevailing method for sequence prediction tasks. The LSTM model can store and retrieve patterns from many sequences. LSTM distinguishes itself from feed-forward neural networks and Recurrent Neural Networks (RNN) by its capacity to utilize past data to generate accurate forecasts[43]. The constituent cells of a conventional LSTM network function as memory units. The initial stage of the procedure involves transmitting the cell's present condition, along with any confidential data. Data is stored in memory blocks, which are discrete units of information. Memory remodeling involves three distinct gates working together. In this study, we investigate the distinct functions performed by individual gates inside a LSTM cell and share our research results. Figure 2 depicts a graphical representation of the different gates. Furthermore, the discourse encompasses the examination of input and output gates, in addition to the forget gate. The initial gate in an LSTM cell is called the Forget Gate. This gate determines whether the previous time step data should be retained or deleted.

The improved LSTM algorithm is able to effectively capture the temporal dynamics of the video, allowing it to recognize more fine-grained motion patterns. The U-Net-based spatial feature extraction method helps the model to capture the overall structure of the video, while the MobileNetV2 encoder component allows it to learn higher-level features from the extracted spatial features. The combination of these two methods allows our model to accurately recognize motion in the video. Analyzing the relationship between the hidden state e_{p-1} and the input variable V_p. The present state of the

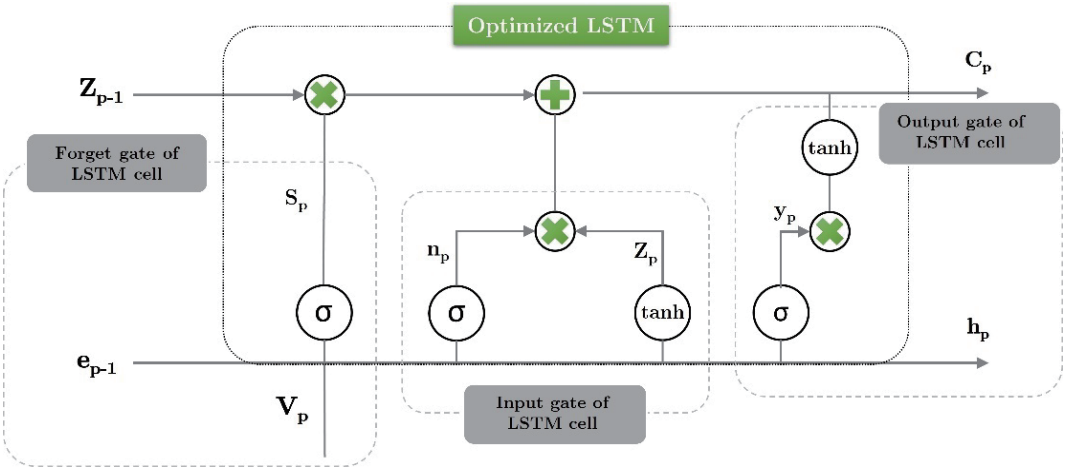

Fig. 2 The classification stage incorporates LSTM to enhance the model's performance, in addition to the previously mentioned layers.

cellular system is multiplied by a sigmoid function whose output varies from 0 to 1 (Eqs. (2)–(6)) in combination with the previous temporal iteration.

If the result is 1, no components are left out. If the result is zero, nothing is kept. The "*D*" variable symbolizes density, whereas the "*g*" variable represents bias.

$$s_p = \sigma\left(D_s \times (e_{p-1}, V_p) + g_s\right) \tag{2}$$

Additionally, both the current state V_p and the preceding present state e_{p-1} within the input gate exhibit the application of the hyperbolic tangent function (tanh). Since a sigmoid function has been used in previous studies, this method seems reasonable. Final value is between 0 and 1, inclusive. Thus, by painstakingly integrating individual values from the input vector n_p, the aforementioned technique ensures the precision of the current cell state *C*.

$$n_p = \sigma\left(D_n \times (e_{p-1}, V_p) + g_n\right) \tag{3}$$

$$Z_p = \tanh\left(D_Z \times (e_{p-1}, V_p) + g_Z\right) \tag{4}$$

At the output gate, a third sigmoid function integrates the data from the visible and invisible states. The changed cell state is subsequently supplied into the tanh function through the Input Gate's output. The hidden state for the following cycle may be calculated by multiplying both outputs by a point value. Both the changed and hidden states are stored for further use.

$$y_p = \sigma\left(D_y \times (e_{p-1}, V_p) + g_y\right) \tag{5}$$

$$e_p = \tanh\left(z_p\right) \times y_p \tag{6}$$

Backpropagation of weights and errors may be accomplished thanks to an LSTM cell's structure. By doing so, we may adjust weights and biases to greatly improve forecast reliability. This tool can also be used to fix gradient amplification and attenuation problems. The modified sequential LSTM model is a multi-layer LSTM, made up of parts 1, 2, and 3. This multi-layered design consists of a data preprocessing layer and a sequence of LSTM layers. In the sequential LSTM model, three layers are LSTMs and the fourth is a dense layer. Though it increases output, adding extra layers is computationally expensive. As a result, four layers offer the most optimal trade-off between efficacy and efficiency. After missing data is filled in, the closest column from the original dataset is utilized as input for the multi-layer structure of the modified sequential LSTM model.

To identify combat movements utilizing the residual grid, temporal data are classified using an LSTM classifier. The cross-entropy loss function (CEL_f) is employed as the error function in this study, since the model exhibits proficiency in performing binary classification tasks pertaining to videos. Presented below is an illustrative instance of an error function employed in model fitting.

$$\mathrm{CEL}_f = \frac{-1}{\mathrm{Output}_{\mathrm{size}}} \times \sum_{i=1}^{\mathrm{Output}_{\mathrm{size}}} \left[(y_i \times \log \mathrm{out}_1) + ((1 - y_i) \times \log \mathrm{out}_2)\right] \tag{7}$$

where,

$$Ou_1 = \log(\hat{y}_1) \tag{8}$$

and,

$$o_2 t_2 = \log(1 - \hat{y}_1) \tag{9}$$

In the given context, let $\hat{y}_I$ denote the *i*-th scalar value in the output of the model, y_i represent the appropriate target value, and output size indicate the overall count of scalar values. The choice of the encoder component in our U-Net-like model for spatial information extraction was made by employing the pre-trained MobileNetV2 algorithm[44]. The decision was made based on the intrinsic qualities of the option.

3.3 Integrated structure

The temporal characteristic is derived by deferring the processing of an auxiliary queue of frame characteristics to the subsequent phase. The optimized LSTM model captures sequential information across consecutive video frames. Two-tier classifiers utilizing dense layers have the capability to differentiate between regular and anomalous crowd behavior as well as art museum activities. The diagram presented in Fig. 3 illustrates the proposed model framework.

In summary, MobileNetV2 showcases the capability of achieving appropriate results while minimizing resource demands, spanning computational and learning parameter elements. Figure 1 illustrates the integration of MobileNet V2 with the U-Net-like feature extractor. The employed model incorporates an encoder that has undergone pre-training using Imagenet data. Training effectiveness can be improved

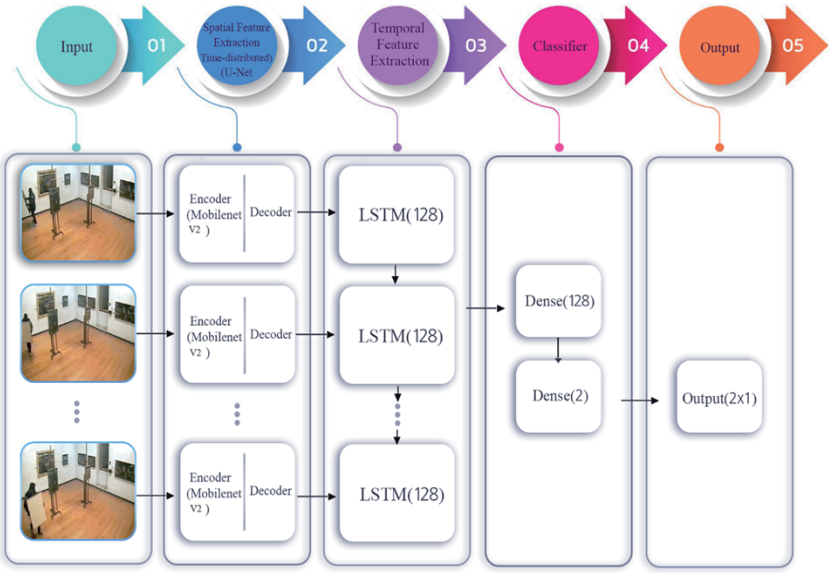

Fig. 3 Proposed architecture that classifies abnormal behavior in video frames received from the art museum.

by including more unlabeled geographical data inside frames. The existing data about crowd occurrences mostly comprises spatiotemporal information, necessitating observation and analysis in a dynamic temporal context rather than static visual representations. Moreover, security camera footage spans a wide array of scenarios. The challenge of training a model to create a relationship between the temporal evolution of certain qualities and crowd events and behaviors becomes less significant when a reliable and efficient spatial-feature extractor is available.

3.4 Optical flow

A tracking system based on optical flow is used after determining abnormal conditions in the art museum environment[45]. As a result of the relative displacement between the object and the camera, optical flow refers to the phenomenon of object mobility between successive frames in a series. A schematic representation of optical flow is shown in Fig. 4.

Within the framework of successive frames, it is feasible to express the luminance of an image (I) as a mathematical function that incorporates both spatial dimensions (x, y) and temporal dimension (t). In essence, the application of a displacement vector (dx, dy) to the pixels of the initial image $I(x, y, t)$ over a temporal interval t results in a transformed image $I(x+\mathrm{d}x, y+\mathrm{d}y, t+\mathrm{d}t)$. The initial assumption is that object pixel intensities stay consistent over consecutive

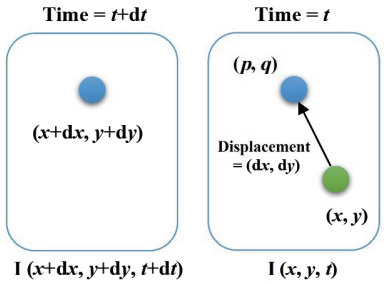

Fig. 4 Optical flow problem is measured by considering the motion of pixels in subsequent frames to monitor the object's motion.

frames.

$$I(x,y,t) = f(x+\delta x, y+\delta y, t+\delta t) \tag{10}$$

Subsequently, the Taylor Series Approximation is employed to estimate the displacement pixel, while concurrently eliminating the frequently encountered components.

$$I(x+\delta x, y+\delta y, t+\delta t) = I(x,y,t) + \frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t \tag{11}$$

where

$$\frac{\partial t}{\partial x}\mathrm{d}x + \frac{\partial t}{\partial y}\partial y + \frac{\partial t}{\partial t}\mathrm{d}t = 0 \tag{12}$$

After that, the optical flow equation is derived by the process of dividing by the temporal interval, dt:

$$\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = 0 \tag{13}$$

Let us examine the aforementioned variables, where

u denotes the derivative of x with respect to t ($\mathrm{d}x/\mathrm{d}t$), and v is the derivative of y with respect to t ($\mathrm{d}y/\mathrm{d}t$). In addition, the derivatives $\mathrm{d}I/\mathrm{d}x$, $\mathrm{d}I/\mathrm{d}y$, and $\mathrm{d}I/\mathrm{d}t$ correspond to the image gradients in relation to the horizontal axis, the vertical axis, and time, respectively. Hence, it may be deduced that the subject under consideration relates to optical flow, which encompasses the determination of $u(\mathrm{d}x/\mathrm{d}t)$ and $v(\mathrm{d}y/\mathrm{d}t)$ to determine motion over a specific duration. It is important to acknowledge that the optical flow equation presents a challenge in terms of solving for the variables u and v. This challenge arises from the fact that there are two unknowns and only one equation in the equation.

4 Result

The evaluation of results is based on the outcomes of implementing the research plan within this context. Our study begins by introducing the collected dataset and examining all relevant criteria within this field.

4.1 Dataset and setting

For this research, various videos collected from YouTube, Vimeo, DTUBE, videos from shutter stock related to the art museum and other searched videos were used. Some specific videos were also used in some datasets such as UCSD, CrowdHuman, UCF-QNRF, Kaggle, and NWPU-Crowd about the population and movement of people in art museums. In total, 58 video collections have been collected, which are more than 24 000 frames related to normal behaviors of people in art museums and 13 000 video frames relevant to abnormal behaviors including crowds, fights, thefts, terror, objects and obstacles, and other harmful behaviors. It was done in relation to valuable objects in the art museum.

The videos had different frame rates from 25 to 30, which were all converted to video frames (images) for manual processing. Each frame was annotated by two experts and it was determined in which frame there is normal and abnormal behavior. The images created as part of the proposed network were 224 × 224 and all frames were in JPG format. Figure 5 shows several video frames that occurred in the art museum. Additionally, before being checked by experts and creating labels for each frame, frames without traffic were also removed from the collection.

4.2 Experimental setup and metrics

The approach described above operates on Windows 10 and is implemented using the MATLAB R2022b programming environment. Regarding the computer hardware, we employed an Intel® CoreTM i5-8500 processor (single central processing unit) accompanied by 16 gigabytes of Random-Access Memory (RAM). Additionally, we incorporated an extra 16 gigabytes of RAM as a contingency measure, which was stored on a Solid-State Drive (SSD). Time distribution in U-Net features extractor includes 30, 64, 64, and 1 output shape and 1 907 041 parameters. LSTM structures have 128 output shapes and 2 163 200 parameters similarly. Moreover, two dense layers have 32 and 2 output shapes, and 4128 and 62 parameters, respectively. The format of all frames has been changed to JPG and the dimensions are equal to 224 × 224. Moreover, we examine the efficacy of the proposed model in accurately classifying and categorizing videos as conventional or unconventional events. The experimental findings confirmed the computational efficiency and rapidity of the developed model, which consists of 4 056 236 parameters. The researchers conducted five-fold cross-validation.

Different recognition systems can be designed depending on the art museum's recognition accuracy.

Fig. 5 A collection of dangerous behaviors in art museums. The first row describes a painting theft. In the second row, the crowd is seen escaping a problem with abnormal behavior.

Comparison can be made between the factor and the labels in the sample data set. It is also possible to display a confusion matrix. Based on the model's performance during the classification phase, True Positives (TPs), False Negatives (FNs), False Positives (FPs), and true negatives are estimated. Aside from classifying the frame based on the combined deep model structure, target estimation is also performed.

A confusion matrix is used in classification. In the section on dividing data to check the whole model's performance, the K-fold method is used to divide data. At this stage, a tracking zone is given to each individual within the input frame. Small frame sizes during tracking reduces computing complexity. Evaluation of a recognition model's efficacy may be assessed by considering its accuracy, recall, and precision. Hence, the calculation of TP, TN, FP, and FN is performed for each iteration of the procedure.

4.3 Evaluations

Table 1 presents the results of the classification section based on the need to declare damage to valuable objects are presented. It has been examined how the combined method affects classification, entry of few people, entry of many people, and entry of people with different numbers (low and high). In this study, we examine the representation of 5-fold cross validation and different data frames. Thus, the proposed method improves classification accuracy by 2 to 4 percent. This table illustrates the various situations of classification accuracy. There is no large dispersion of responses due to the low share of changes. As a result, the method is highly robust to various frame changes. Based on the average classification accuracy of 97.67%, the algorithm creates a suitable monitoring response. Figure 6 displays the confusion matrix for classification in three modes of low, medium, and high complexity. The average accuracy loss is about 1.3%, which is not considered significant compared to other similar methods.

However, the proposed method for classifying

Table 1 A proposed model identifies abnormal behavior of crowds or humans in art museums based on the results of the classification section. Based on the classification of video frames into three complexity states, low, high, and medium, the proposed model is compared with similar methods in this field.

(%)

Model	Museum environment with low crowd density			Museum environment with moderate crowd density			Museum environment with high crowd density		
	U-Net	U-Net with MobileNet V2	U-Net with MobileNet V2 and optLSTM	U-Net	U-Net with MobileNet V2	U-Net with MobileNet V2 and optLSTM	U-Net	U-Net with MobileNet V2	U-Net with MobileNet V2 and optLSTM
5-fold (1)	95.24	96.89	97.21	94.78	96.28	96.91	94.44	96.03	96.47
5-fold (2)	95.61	96.98	97.42	95.02	96.59	97.01	94.56	96.19	96.58
5-fold (3)	95.13	96.80	97.14	94.81	96.22	96.80	94.21	95.90	96.29
5-fold (4)	95.24	96.77	97.06	94.56	96.15	96.75	94.11	95.83	96.12
5-fold (5)	95.42	96.52	97.30	94.60	96.66	97.03	93.75	96.12	96.56

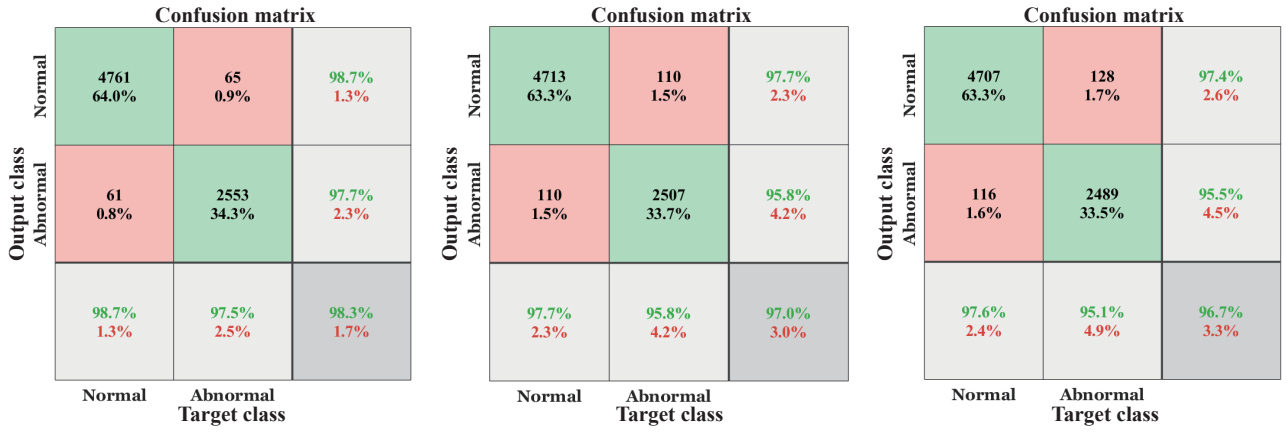

Fig. 6 Classification section confusion matrix example. From left to right, the classification estimation and feature extraction sections for museum environments are shown for low, medium, and high crowd complexity.

normal and abnormal behavior in crowds visiting different art museum environments under different lighting conditions has effectively analyzed the different conditions of receiving video frames.

Table 2 presents a succinct overview of the tracking approach's effectiveness across various audience sizes. This is within the context of both large and small frame sizes commonly encountered in museum settings. Using this table, the process of tracking people in art museums was examined three times by classifying the frames into three categories: low density of crowd, moderate density of crowd, and high density of crowd.

Additionally, for the optical flow method of identification, the correct and incorrect tracking were automatically calculated, as well as compared to the ground truth frames that were manually separated. As the quality of the images decreases (especially when the frames have a low resolution), the algorithm still provides a satisfactory response, even after repeated tests. Furthermore, accuracy, sensitivity, and specificity have also been affected by low, medium, and high crowds. The method of optical flow and its output have been demonstrated in a series of videos in Fig. 7.

Some factors, such as shadows of people or the angle of the camera, affect accuracy, such as shadows of people. Although the proposed deep structure has significantly improved feature extraction optical flow, its accuracy can be improved by more optimal settings. A set of frames with a specified optics direction can be used to investigate instantaneous movement states.

The frames with a specific optics direction can help to identify any changes in the optical flow and provide a more accurate estimation of the object's motion. This can be used to develop an algorithm that can more accurately recognize the motion of objects in the environment. Additionally, the use of deep structures in the feature extraction process can help reduce the impact of shadows and other external factors. The output of optical flow can be used to provide an estimate of the speed of an object, as well as the direction of its motion. This can be used to automate the tracking of objects in the environment, as well as to help recognize potential threats. The main difference between optical flow-based tracking and other comparable methods such as Gaussian mixture models and Kalman filters is this problem. Figure 7 shows the tracking algorithm performs well under different conditions with few and many people in the videos.

Moreover, environmental images face some challenges, such as excessive ambient light intensity, clothes similar to the environment, too much inactivity in the frames, too much darkness in the environment and similar things that can be overcome by some processes. Tracking has, however, been remarkable with the algorithm.

5 Discussion

Using video frames collected from normal and abnormal crowd behavior in several art museum environments, the proposed method involves classification and tracking procedures. As a result, guardians have the opportunity to preserve expensive objects by identifying abnormal events in their environments correctly and accurately. An improved deep structure algorithm combining U-Net with MobileNetV2 and optLSTM architectures analyzes and

Table 2 Three times, the process of tracking people in art museums was examined by manually dividing the frames into three categories: low density of crowd, median density of crowds, and high density of crowds. Furthermore, three experiments were conducted and the accuracy (Acc), sensitivity (Sen), and specificity (Spe) of video frames of low, moderate, and high quality were evaluated.

Experiment	Dataset (frame quality)	Low density of crowd			Median density of crowd			High density of crowd		
		Acc	Sen	Spe	Acc	Sen	Spe	Acc	Sen	Spe
Exp 1	Low	0.9811	0.9789	0.9832	0.9711	0.9642	0.9718	0.9692	0.9620	0.9699
	Moderate	0.9855	0.9817	0.9890	0.9767	0.9730	0.9826	0.9731	0.9680	0.9766
	High	0.9881	0.9828	0.9914	0.9785	0.9755	0.9913	0.9742	0.9722	0.9801
Exp 2	Low	0.9766	0.9751	0.9772	0.9698	0.9635	0.9709	0.9680	0.9632	0.9702
	Moderate	0.9786	0.9721	0.9796	0.9744	0.9718	0.9791	0.9703	0.9673	0.9714
	High	0.9829	0.9808	0.9891	0.9724	0.9712	0.9860	0.9708	0.9695	0.9764
Exp 3	Low	0.9830	0.9794	0.9844	0.9740	0.9722	0.9781	0.9702	0.9652	0.9712
	Moderate	0.9843	0.9829	0.9911	0.9790	0.9752	0.9824	0.9742	0.9716	0.9776
	High	0.9863	0.9840	0.9922	0.9764	0.9750	0.9876	0.9732	0.9723	0.9804

Fig. 7 Tracking algorithm was capable of showing the presence of people in different closed environments of the art museum, which tracked the movements of people. An analyzed frame is selected every 20 frames.

classifies video frames for different purposes. An in-depth analysis of our thinking and findings will follow, followed by a detailed description of each method.

The evaluation of a test's overall effectiveness in comparison to other tests may be effectively conducted by utilizing the area under the Receiver Operating Characteristic curve (ROC) as a valuable statistic. This evaluation can be conducted alongside a comparison of alternative methodologies. Greater values in this statistical measure are statistically associated with a higher degree of realism. Figure 8 illustrates several model identification techniques, represented by Area Under the Curve (AUC) and ROC curves. Considerable attention has been devoted to examining the robustness of ROC curves within the framework of using received video frames. This is for recognizing anomalous occurrences in art museum environments.

The presence of anomalies is a potential vulnerability at art museums' locations. Various factors, including ambient noise, available illumination, disparities in

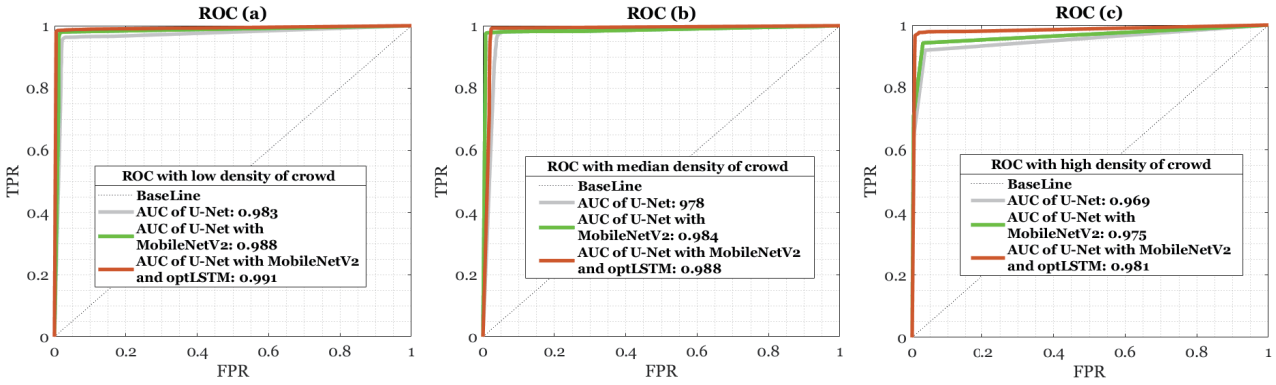

Fig. 8 Comparison of the proposed method with two similar models. Comparing methods based on (a) low complexity frames, (b) medium complexity frames, and (c) high complexity frames.

subject size, and camera position, have a significant influence. To conduct a comprehensive analysis, it is necessary to assess the contextual information acquired from the video frames. ROC curve analysis has been established as a reliable method for evaluating the effects of unforeseen occurrences in art museum environments.

According to the study findings, the ROC curve has the potential to be utilized as a predictive tool for recognizeing atypical occurrences in art museum environments. Figure 8 illustrates the implemented methodology and the corresponding class representations. In bustling art museum environments, these categories were shown separately to demonstrate both typical and unconventional human conduct. The measurement known as the AUC has been employed as a statistical tool for evaluating and contrasting various methodologies aimed at determining the occurrence rate of atypical occurrences in art museum environments. These occurrences are often categorized as low, medium, or high crowd complexity.

The utilization of AUC provides a quantitative measure to evaluate the overall performance of the predictive model across varying levels of crowd complexity, contributing valuable insights for enhancing security and management strategies in art museum settings.

5.1 Comparison

This study evaluates the performance of several architectures in terms of execution speed for various transfer learning algorithms. Furthermore, it is hypothesized that the proposed approach may attain the desired level of accuracy in execution and computational efficiency by employing U-Net with MobileNetV2 and optLSTM, despite its limited feature set. Hence, this approach proves advantageous in several scenarios, including embedded systems and real-time software implementations. The duration of processing time has been significantly influenced by the extensive feature generation procedure. Nevertheless, the system maintains precision while operating in real-time, a feat made achievable by a judicious balance between accuracy and computational complexity. The proposed strategy's viability is enhanced by its scalability potential. Accurate feature extraction plays a critical role in maximizing classification performance and minimizing computational costs associated with all classification algorithms. The approach described above has demonstrated its accuracy and resilience via empirical data, even when utilizing a limited collection of variables.

The current body of literature about video processing and monitoring effectiveness in crowd

behavior analysis, particularly in relation to art museum visitors, is limited. This scarcity of study poses challenges to comprehensively examining and understanding abnormal human behavior in such settings. One notable obstacle is the absence of a universally accepted protocol for acquiring a dataset valuable for research purposes. Numerous methodologies have been proposed for the investigation of this particular field of crowd analysis. One of the ways commonly discussed in the literature is the non-automatic approach to assessing abnormal crowd behavior in art museums, which involves a limited number of steps. The findings derived from our study pertaining to the classification, tracking, and monitoring of individuals inside ordinary environments

captured in video frames may be juxtaposed with prior research conducted in this domain. Hence, our model exhibits a higher level of accuracy than the current leading options in the field. Efficiency, robustness, and dependability of the model are all noteworthy. Hence, it may be relied upon to accurately assess human motion and deviant actions, thereby assisting in the deployment of appropriate remedial measures. To evaluate the efficacy of the proposed methodology in decision-making, we have included a comparative analysis of established techniques in Table 3. Although these methods are not about classification and tracking crowd analysis and behavior in art museums, we compared them and the proposed method to justify the method.

We considered it significant to explain why the optimized LSTM architecture in their research performs better than the conventional LSTM architecture. Hence, we introduced enhancements to improve the model's ability to capture and process temporal features in video surveillance data. We achieved improved performance in identifying abnormal behavior and enhancing surveillance in art museums compared to the standard LSTM architecture.

When it comes to spotting irregular behavior in large art museum crowds, our model outperforms state-of-the-art approaches. In this study, we compared several models and empirical results. While the suggested model is effective at picking out abnormal humans in large groups, it might need some accuracy work. The number of model parameters can be reduced, which is an additional option. Future studies might benefit from using real-world datasets, instead of museum-only data. This is because museum-collected video stills may not adequately represent regular audience behavior. There is a lot of room for growth in this field of study. This might lead to a better understanding of the underlying activity patterns that influence crowd behavior. Moreover, LSTM classifiers can recognize complex patterns in videos, such as facial expressions, gestures, and body language. By combining an LSTM classifier with a locality-sensitive algorithm, we can build trust-aware LSTM models that take into account both the content and the context of the videos[51].

5.2 Limitation and opportunity

When deep learning accuracy is affected by noise sources, such as environmental noise and imaging devices, several consequences may arise in the context of analyzing video surveillance data from art museums:

1. Impact on real-time processing: Noise introduces inconsistencies and distortions in video data, which can hinder hybrid edge-to-cloud servers' ability to process live videos in real-time. Processing time and resource requirements may increase due to noisy input.

2. Decreased accuracy in abnormal behavior detection: Noise can negatively affect abnormal behavior detection in art museums. The presence of environmental noise or variations in imaging devices can introduce additional patterns or variations in the data. This can lead to false positives or false negatives in abnormal behavior detection.

3. Reduced feature extraction effectiveness: Spatial and temporal feature extraction methods, such as those based on the U-Net architecture, MobileNetV2, and LSTM, may be compromised by noise. Noise can disrupt the extraction of relevant features, making it harder to capture and represent meaningful patterns related to crowd motion or individual behavior.

4. Compromised reliability and generalizability: The proposed methodology for analyzing atypical museum visitor behavior may suffer from reduced reliability and generalizability in noise. Noise can introduce inconsistencies in the training data, leading to models that are overly sensitive or biased towards noise present during training. This limits their ability to generalize to unseen data or different museum environments.

Table 3 Major parameters we use to determine crowd analysis success in video frames are accuracy and computing cost. We evaluate the results obtained using our suggested strategy compared to those obtained through alternative methods.

Reference	Model	Accuracy	Computational cost
Zhou et al.[46]	AnomalyNet	94.4%	High
Rezaee et al.[47]	Modified ResNet architecture	96.55%	Moderate
Li et al.[48]	ST-CaAE	95.5%	Moderate
Chu et al.[49]	SCG-SF	90.9%	Moderate
Singh et al.[50]	Aggregation of ensembles	95.25%	High
Proposed architecture	U-Net with MobileNetV2 and optLSTM	97.67%	Low

5. Robustness challenges: Addressing noise sources becomes critical to ensure the robustness of the deep learning model for video analysis in art museums. Robustness techniques, such as data preprocessing, denoising, or data augmentation, may need to be applied to mitigate the impact of noise. This will improve the model's ability to handle noisy input and adapt to different noise conditions.

The ability to recognize crowd behavior in videos is crucial for a wide variety of uses. When applied to large datasets of video frames, the suggested approach shows impressive accuracy in identifying crowd activity. There's a lot happening in the scene's middle, but things slow down significantly after that. Due to the lack of random occurrences, the high quality of the recordings, and the center location of the event inside the visual frame, artificial circumstances are generated to discover anomalous crowd behavior in art museums. Real-world movies rarely depict human animosity. This event can occur anywhere within the human visual field and last from a fraction of a second to many minutes. Moreover, deep networks include those used for recognizing crowd behavior, creating algorithms for flying drones, employing the Internet of Things (IoT) in video processing[52, 53], cloud computing, edge computing, and classifying data from situations with sparse annotations.

6 Conclusion

Art museum security issues span a wide range, ranging from a variety of threats like armed robberies to more dynamic ones like overcrowding. This necessitates the ability to identify a wide variety of abnormal behaviors. Therefore, it is argued that there is considerable room for improvement in the suggested approach to identifying aberrant conduct in contemporary crowds. Therefore, in this study, we provide a novel and efficient method for recognizing suspicious behavior in art museum surveillance footage. We combine techniques from video processing and deep learning. We propose a U-Net-like network based on MobileNetV2 as the encoder and an enhanced LSTM for temporal feature extraction and classification as a model for spatial feature extraction. This model's low resource needs are the direct result of its careful construction. We used a huge dataset of video frames kindly provided by art institutions and performed a 5-fold cross-validation method. In experiments using complex video frames from real security cameras, the accuracy was 97.67 + 1.23%. The proposed model achieved high accuracy while still being relatively easy to implement and computationally cheap. Our idea works well in time-sensitive environments and on edge devices. In order to enhance the accuracy and robustness of abnormal behavior recognition in art museums, we will expand datasets, integrate, and fusion multiple deep learning techniques. Furthermore, it is critical to develop a model that is capable of detecting out-of-the-ordinary behavior in a variety of environments, camera configurations, lighting variations, and crowd dynamics common in art museums. Future works include improving model efficiency, enabling real-time analysis on edge devices or in time-sensitive environments, and leveraging advancements in deep learning.

References

[1] G. Sreenu and M. A. Saleem Durai, Intelligent video surveillance: A review through deep learning techniques for crowd analysis, *J. Big Data*, vol. 6, no. 1, p. 48, 2019.

[2] J. Laufs, H. Borrion, and B. Bradford, Security and the smart city: A systematic review, *Sustain. Cities Soc.*, vol. 55, p. 102023, 2020.

[3] L. Fei and B. Han, Multi-object multi-camera tracking based on deep learning for intelligent transportation: A review, *Sensors*, vol. 23, no. 8, pp. 3852, 2023.

[4] T. I. Amosa, P. Sebastian, L. I. Izhar, O. Ibrahim, L. S. Ayinla, A. A. Bahashwan, A. Bala, and Y. A. Samaila, Multi-camera multi-object tracking: A review of current trends and future advances, *Neurocomputing*, vol. 552, p. 126558, 2023.

[5] N. Bisagno, N. Conci, and B. Zhang, Data-driven crowd simulation, in *Proc. 14th IEEE Int. Conf. Advanced Video and Signal Based Surveillance* (*AVSS*), Lecce, Italy, 2017, pp. 1–6.

[6] Z. Zhang, X. Wang, D. Huang, X. Fang, M. Zhou, and B. Mi, iDT: An integration of detection and tracking toward low-observable multipedestrian for urban autonomous driving, *IEEE Trans. Ind. Inf.*, vol. 19, no. 9, pp. 9887–9897, 2023.

[7] F. Camara, N. Bellotto, S. Cosar, D. Nathanael, M. Althoff, J. Wu, J. Ruenz, A. Dietrich, and C. W. Fox, Pedestrian models for autonomous driving Part I: Low-level models, from sensing to tracking, *IEEE Trans. Intell. Transport. Syst.*, vol. 22, no. 10, pp. 6131–6151, 2021.

[8] S. R. Runhovde, The art of balancing accessibility and security in museums, *J. Risk Res.*, vol. 24, no. 9, pp. 1113–1126, 2021.

[9] Z. Huo, Legal protection of cultural heritage in China: A challenge to keep history alive, *Int. J. Cult. Policy*, vol. 22, no. 4, pp. 497–515, 2016.

[10] M. G. Rashed, R. Suzuki, T. Yonezawa, A. Lam, Y.

Kobayashi, and Y. Kuno, Tracking visitors in a real museum for behavioral analysis, in *Proc. Joint 8th Int. Conf. Soft Computing and Intelligent Systems* (*SCIS*) *and 17th Int. Symp. on Advanced Intelligent Systems* (*ISIS*), Sapporo, Japan, 2016, pp. 80–85.

[11] J. M. Grant and P. J. Flynn, Crowd scene understanding from video: A survey, *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 13, no. 2, p. 19, 2017.

[12] N. Wojke, A. Bewley, and D. Paulus, Simple online and realtime tracking with a deep association metric, in *Proc. IEEE Int. Conf. Image Processing* (*ICIP*), Beijing, China, 2017, pp. 3645–3649.

[13] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu, and N. Yu, Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism, in *Proc. IEEE Int. Conf. Computer Vision* (*ICCV*), Venice, Italy, 2017, pp. 4836–4845.

[14] H.-M. Hsu, T.-W. Huang, G. Wang, J. Cai, Z. Lei, and J.-N. Hwang, Multi-camera tracking of vehicles based on deep features re-ID and trajectory-based camera link models, https://www.semanticscholar.org/paper/Multi-Camera-Tracking-of-Vehicles-based-on-Deep-and-Hsu-Huang/64366fd28a9cecdf3156b1fa84d1613ec8161f1c, 2019.

[15] N. M. Al-Shakarji, F. Bunyak, G. Seetharaman, and K. Palaniappan, Multi-object tracking cascade with multi-step data association and occlusion handling, in *Proc. 15th IEEE Int. Conf. Advanced Video and Signal Based Surveillance* (*AVSS*), Auckland, New Zealand, 2018, pp. 1–6.

[16] M. Rainoldi, C.-E. Yu, and B. Neuhofer, The museum learning experience through the visitors' eyes: An eye tracking exploration of the physical context, in *Eye Tracking in Tourism*, M. Rainoldi and M. Jooss, eds. Cham, Switzerland: Springer, 2020, pp. 183–199.

[17] J. A. Donaire, N. Galí, and B. Gulisova, Tracking visitors in crowded spaces using zenith images: Drones and time-lapse, *Tour. Manag. Perspect.*, vol. 35, p. 100680, 2020.

[18] B. Gaikwad and A. Karmakar, Smart surveillance system for real-time multi-person multi-camera tracking at the edge, *J. Real Time Image Process.*, vol. 18, no. 6, pp. 1993–2007, 2021.

[19] Z. Shao, J. Cai, and Z. Wang, Smart monitoring cameras driven intelligent processing to big surveillance video data, *IEEE Trans. Big Data*, vol. 4, no. 1, pp. 105–116, 2018.

[20] T. Wang, M. Z. A. Bhuiyan, G. Wang, L. Qi, J. Wu, and T. Hayajneh, Preserving balance between privacy and data integrity in edge-assisted Internet of Things, *IEEE Internet Things J.*, vol. 7, no. 4, pp. 2679–2689, 2020.

[21] F. Wang, L. Wang, G. Li, Y. Wang, C. Lv, and L. Qi, Edge-cloud-enabled matrix factorization for diversified APIs recommendation in mashup creation, *World Wide Web*, vol. 25, no. 5, pp. 1809–1829, 2022.

[22] P. Centorrino, A. Corbetta, E. Cristiani, and E. Onofri, Managing crowded museums: Visitors flow measurement, analysis, modeling, and optimization, *J. Comput. Sci.*, vol. 53, p. 101357, 2021.

[23] A. Ferrato, C. Limongelli, M. Mezzini, and G. Sansonetti, Using deep learning for collecting data about museum visitor behavior, *Appl. Sci.*, vol. 12, no. 2, p. 533, 2022.

[24] J. Lanir, T. Kuflik, J. Sheidin, N. Yavin, K. Leiderman, and M. Segal, Visualizing museum visitors' behavior: Where do they go and what do they do there, *Pers. Ubiquitous Comput.*, vol. 21, no. 2, pp. 313–326, 2017.

[25] L. Seidenari, C. Baecchi, T. Uricchio, A. Ferracani, M. Bertini, and A. Del Bimbo, Deep artwork detection and retrieval for automatic context-aware audio guides, *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 13, no. 3s, pp. 1–21, 2017.

[26] M. Mezzini, C. Limongelli, G. Sansonetti, and C. De Medio, Tracking museum visitors through convolutional object detectors, in *Proc. Adjunct Publication of the 28th ACM Conf. on User Modeling, Adaptation and Personalization* (*UMAP '20 Adjunct*), Genoa, Italy, 2020, pp. 352–355.

[27] K. Trejo, C. Angulo, S. Satoh, and M. Bono, Towards robots reasoning about group behavior of museum visitors: Leader detection and group tracking, *J. Ambient Intell. Smart Environ.*, vol. 10, no. 1, pp. 3–19, 2018.

[28] P. Centorrino, A. Corbetta, E. Cristiani, and E. Onofri, Measurement and analysis of visitors' trajectories in crowded museums, arXiv preprint arXiv:1912.02744, 2019.

[29] C. Martella, A. Miraglia, J. Frost, M. Cattani, and M. van Steen, Visualizing, clustering, and predicting the behavior of museum visitors, *Pervasive Mob. Comput.*, vol. 38, pp. 430–443, 2017.

[30] F. Zafari, A. Gkelias, and K. K. Leung, A survey of indoor localization systems and technologies, *IEEE Commun. Surv. Tutorials*, vol. 21, no. 3, pp. 2568–2599, 2019.

[31] M. Fiorucci, M. Khoroshiltseva, M. Pontil, A. Traviglia, A. Del Bue, and S. James, Machine learning for cultural heritage: A survey, *Pattern Recognit. Lett.*, vol. 133, pp. 102–108, 2020.

[32] A. Augello, I. Infantino, G. Pilato, and G. Vitale, Site experience enhancement and perspective in cultural heritage fruition—A survey on new technologies and methodologies based on a four-pillars approach, *Future Internet*, vol. 13, no. 4, p. 92, 2021.

[33] P. Roy and C. Chowdhury, A survey of machine learning techniques for indoor localization and navigation systems, *J. Intell. Rob. Syst.*, vol. 101, no. 3, p. 63, 2021.

[34] Y. Desmarais, D. Mottet, P. Slangen, and P. Montesinos, A review of 3D human pose estimation algorithms for markerless motion capture, *Comput. Vis. Image Underst.*, vol. 212, p. 103275, 2021.

[35] S. Hong, T. Yi, J. Yum, and J.-H. Lee, Visitor-artwork network analysis using object detection with image-retrieval technique, *Adv. Eng. Inform.*, vol. 48, p. 101307, 2021.

[36] N. Saito, F. Kusunoki, S. Inagaki, and H. Mizoguchi, Novel application of an RGB-D camera for face-direction measurements and object detection: Towards understanding museum visitors' experiences, in *Proc. 13th Int. Conf. Sensing Technology* (*ICST*), Sydney, Australia,

2019, pp. 1–4.
[37] R. Angeloni, R. Pierdicca, A. Mancini, M. Paolanti, and A. Tonelli, Measuring and evaluating visitors' behaviors inside museums: the Co. ME. project, *SCIRES-ITSCIentific RESearch and Information Technology*, vol. 11, no. 1, pp. 167–178, 2021.
[38] P. Centorrino, A. Corbetta, E. Cristiani, and E. Onofri, Managing crowded museums: Visitors flow measurement, analysis, modeling, and optimization, *J. Comput. Sci.*, vol. 53, p. 101357, 2021.
[39] N. Wang, X. Zhao, P. Zhao, Y. Zhang, Z. Zou, and J. Ou, Automatic damage detection of historic masonry buildings based on mobile deep learning, *Automation in Construction*, vol. 103, pp. 53–66, 2019.
[40] A. Ferrato, C. Limongelli, M. Mezzini, and G. Sansonetti, The META4RS Proposal: Museum Emotion and Tracking Analysis For Recommender Systems, in *Adjunct Proc. of the 30th ACM Conf. on User Modeling, Adaptation and Personalization*, New York, NY, USA, 2022, pp. 406–409.
[41] T. Yi, H. Kim, and J.-H. Lee, Predicting museum visitors' artwork preference through deep learning, *Arch. Des. Res.*, vol. 35, no. 4, pp. 309–323, 2022.
[42] X. Yang and J. A. Esquivel, Time-aware LSTM neural networks for dynamic personalized recommendation on business intelligence, *Tsinghua Science and Technology*, vol. 29, no. 1, pp. 185–196, 2024.
[43] X. Yang and J. A. Esquivel, LSTM network-based adaptation approach for dynamic integration in intelligent end-edge-cloud systems, *Tsinghua Science and Technology*, vol. 29, no. 4, pp. 1219–1231, 2024.
[44] M. Elgendi, M. U. Nasir, Q. Tang, R. R. Fletcher, N. Howard, C. Menon, R. Ward, W. Parker, and S. Nicolaou, The performance of deep neural networks in differentiating chest X-rays of COVID-19 patients from other bacterial and viral pneumonias, *Front. Med.*, vol. 7, p. 550, 2020.
[45] K. Rezaee, S. M. Rezakhani, M. R. Khosravi, and M. K. Moghimi, A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance, *Pers. Ubiquitous Comput.*, vol. 28, no. 1, pp. 135–151, 2024.
[46] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, and R. S. M. Goh, AnomalyNet: An anomaly detection network for video surveillance, *IEEE Trans. Inform. Forensic Secur.*, vol. 14, no. 10, pp. 2537–2550, 2019.
[47] K. Rezaee, H. G. Zadeh, C. Chakraborty, M. R. Khosravi, and G. Jeon, Smart visual sensing for overcrowding in COVID-19 infected cities using modified deep transfer learning, *IEEE Trans. Ind. Inf.*, vol. 19, no. 1, pp. 813–820, 2023.
[48] N. Li, F. Chang, and C. Liu, Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes, *IEEE Trans. Multimedia*, vol. 23, pp. 203–215, 2021.
[49] W. Chu, H. Xue, C. Yao and D. Cai, Sparse coding guided spatiotemporal feature learning for abnormal event detection in large videos, *IEEE T. Multimedia*, vol. 21, no. 1, pp. 246–255, 2019.
[50] K. Singh, S. Rajora, D. K. Vishwakarma, G. Tripathi, S. Kumar, and G. S. Walia, Crowd anomaly detection using aggregation of ensembles of fine-tuned ConvNets, *Neurocomputing*, vol. 371, no. C, pp. 188–198, 2020.
[51] D. Li and J. A. Esquivel, Trust-aware hybrid collaborative recommendation with locality-sensitive hashing, *Tsinghua Science and Technology*, 2023.
[52] W. Dou, X. Zhao, X. Yin, H. Wang, Y. Luo, and L. Qi, Edge computing-enabled deep learning for real-time video optimization in IIoT, *IEEE Trans. Ind. Inf.*, vol. 17, no. 4, pp. 2842–2851, 2021.
[53] M. R. Khosravi and S. Samadi, Mobile multimedia computing in cyber-physical surveillance services through UAV-borne Video-SAR: A taxonomy of intelligent data processing for IoMT-enabled radar sensor networks, *Tsinghua Science and Technology*, vol. 27, no. 2, pp. 288–302, 2022.

Rongju Yao received the bachelor's degree in fine arts from Jilin University of the Arts in 2007, the master's degree in industrial design engineering from Qingdao University in 2013, and the PhD degree in art education from University of Perpetual Help System DALTA in 2023. She is currently a faculty of Weifang University of Science & Technology. Her current main research direction is digital media.

Mukun Wang received the bachelor's degree in painting from Nanjing University of the Arts in 2006, the master's degree in fine arts (oil painting) from Nanjing University of the Arts in 2010, and is currently pursuing the PhD degree in design (Service design) from Dongseo University in Korea. Her research focuses on the intersection of visual identity technology and Service design.

Khosro Rezaee received the MSc and PhD degrees in biomedical engineering from Hakim Sabzevari University, Sabzevar, Iran, in 2014 and 2018, respectively. His current position is assistant professor at Meybod University in Meybod, Iran. Among his research interests are machine learning, medical image analysis, healthcare systems, deep learning, signal processing, and evolutionary computation.