Journal of Theoretical and Applied Information Technology
31st March 2025. Vol.103. No.6

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EXPLORING THE ADVANCEMENTS AND CHALLENGES OF OBJECT DETECTION IN VIDEO SURVEILLANCE THROUGH DEEP LEARNING: A SYSTEMATIC LITERATURE REVIEW AND OUTLOOK

M.KOTESWARA RAO[1,2], P.M. ASHOK KUMAR2[3]

[1]Research Scholar, Koneru Lakshmaiah Education Foundation, Department of Computer Science Engineering, Vaddeswaram, AP. India.
[2]Assistant Professor, VNR Vignana Jyothi Institute of Engineering and Technology, Department of Information Technology, Hyderabad, India.
[3]Associate Professor, Koneru Lakshmaiah Education Foundation, Department of Computer Science Engineering, Vaddeswaram, AP. India
E-mail: [1]mailtomkrao@gmail.com, [2]pmashokk@gmail.com

ABSTRACT

Video surveillance has become increasingly important in recent years, leading to a growing need for effective and accurate surveillance systems that can detect objects and understand scenes. To address these challenges, methods based on deep learning have emerged as the state-of-the-art approach and have shown remarkable results in various applications. In this systematic literature review (SLR) paper, we investigate the recent advancements and challenges of object detection in video surveillance through deep learning. Our primary goal aim to present a comprehensive overview of the recent research developments in this field and to emphasize the challenges that should be addressed in future research. In this study, we analyzed various deep learning-based object detection methods and evaluated their performances based on several performance metrics. Our findings indicate that deep learning-based methods have demonstrated promising results in regards of accuracy and real-time performance in video surveillance for object detection. However, the study also highlights several challenges such as scalability, robustness, and interpretability that require further research. Finally, this SLR paper concludes with a discussion of future research directions in this field and offers a roadmap for future work. Our study results can serve as a useful reference for researchers and practitioners working in the field of video surveillance and deep learning.

Keywords: *Object detection, Video surveillance, Deep learning, Object tracking, Scene understanding, Performance analysis*

1. INTRODUCTION

Real-time object detection is a basic research direction in the fields of computer vision, deep learning, artificial intelligence. Object detection is essential in video surveillance for improving security, monitoring the movement of objects, and offering valuable insights. It is an important prerequisite for complex computer vision tasks, such as target tracking, event detection and scene understanding. The conventional object detection algorithms manually focus on extracting features include stages like preprocessing, window sliding, feature extraction, feature selection, feature classification and postprocessing. The disadvantages of these techniques mainly include small data size, less portability, high time complexity, redundancy for variety of changes. The emergence of deep learning methods has significantly enhanced the speed and accuracy of object detection in videos. According to recent research in video surveillance object detection has many practical applications, including identifying individuals, detecting anomalies, and tracking vehicles.

Object detection in video surveillance is crucial for enhancing public safety, preventing crimes, and improving traffic management [1]. It is further stated that object detection in video surveillance can be used to find and locate humans and vehicles, and other objects in real-time. The system can then alert security personnel to potential threats, allowing them to respond quickly to potential threats. Furthermore,

object detection can be used to monitor large crowds and detect unusual behaviors, such as pickpocketing or theft. Thus, it is evident that as deep learning algorithms evolve, object detection in video surveillance has become a promising solution to improve public safety and security.

The use of deep learning techniques has transformed the field of object detection in video surveillance, improving the efficiency and accuracy of security systems. Deep learning algorithms provide an unprecedented level of accuracy and speed, allowing for the automatic detection and classification of objects of interest in real-time, reducing the number of false alarms generated by traditional surveillance systems. In addition to improving the accuracy and efficiency of security systems, deep learning has also proven to be an effective method of enhancing the functionality of existing surveillance systems. It can be used to enhance tracking capabilities, even in crowded environments where objects may be occluded or difficult to track. Deep learning can also analyze and extract valuable information from large amounts of surveillance data, enabling security personnel to make more informed decisions and respond more effectively to potential threats.

One of the most commonly used and highly favored deep learning methods for detecting objects in video surveillance is the Convolutional Neural Network (CNN). CNNs are made to learn object hierarchy representations by analyzing the correlations between pixels and convolving the input image to produce feature maps [2]. CNNs can increase the accuracy and speed of object detection in video surveillance. Luo *et al.* [3] presented an instance of utilizing CNNs in the domain of object detection in video surveillance. Their proposed framework utilizes two-stage CNNs to improve object detection and classification precision. They utilized the Faster R-CNN algorithm, which incorporates both object classification a region proposal to achieve exceptional accuracy.

Recurrence Neural Networks (RNNs) are another deep learning method that has been utilized in object detection within video surveillance. RNNs are specifically designed to analyze sequential data, such as video, and are capable of handling the temporal dimension of such data [4]. Object detection in video surveillance can benefit from RNNs by utilizing the temporal information to achieve better results. Haque *et al.* [2] also demonstrated the application of RNNs in object detection for video surveillance. In their proposed methodology, they integrated a long short-term memory network to analyze sequential data and improve object detection accuracy in real-time scenarios. Results indicated that the proposed RNN-based object detection method outperformed traditional object detection methods in both accuracy and computational efficiency.

Deep learning technique YOLO (You Only Look Once) has also been employed in object detection for video surveillance, particularly in the context of crowd density analysis. Salagrama *et al.* [5] utilized YOLO to identify and track individuals in real-time video streams captured by cameras placed in public locations. The authors found that YOLO's real-time processing capabilities and high accuracy in detecting objects made it suitable for crowd density analysis, which can aid in crowd management and emergency response planning.

Figure.1 Object Detection Using YOLO [76]

Another application of YOLO in video surveillance is for security and surveillance purposes. In a study by Saponara *et al.* [6], in real-time video streams captured by security cameras YOLO was applied for object detection and object classification. The authors found that YOLO outperformed other object detection techniques in regards of both speed and accuracy, making it a promising tool for security and surveillance applications. The results indicate that YOLO has the capability to improve the efficiency and effectiveness of video surveillance systems, especially in high-security environments where real-time object detection is critical.

YOLO is a popular framework for real-time to detect object that utilizes a single CNN for the task. The framework is recognized for its rapid inference speed and high accuracy. N-YOLO is a variant of YOLO that improves its efficiency by using multiple smaller networks. R-CNN (Regions with CNN features) is a family of object detection algorithms that utilizes a combination of region proposals and CNN to perform object detection. Ghatwary *et al.* [7] proposed another notable algorithm called Single Shot MultiBox Detector for object detection. This algorithm utilizes a single deep neural network to predict object locations and class probabilities in a single forward pass of the network. To evaluate the performance of object detection algorithms based on deep learning, the Common Objects in Context (COCO) dataset is a widely used large-scale benchmark dataset for object detection and segmentation.

2. AIM OF STUDY

The primary objective of this study is to compare and evaluate novel deep learning-based object detection models that address the limitations of existing approaches in real-world applications. This study aims to enhance accuracy, generalizability, and computational efficiency by integrating advanced architectures, hybrid models, and domain-specific adaptations. Unlike prior studies, this review focuses on optimizing detection in diverse conditions, such as low-light environments, occlusions, and real-time surveillance scenarios.

This study aims to contribute to the ongoing research by providing a well-structured comparative analysis of existing models and proposing solutions that enhance the effectiveness of deep learning-based object detection in video surveillance.

3. RELATED WORK

There are numerous papers have done based on this research topics. Here are some papers that survey related to object detection in deep learning are:

Zou *et al.* [8] explored a range of crucial topics, such as milestone detectors, detection datasets, evaluation metrics, fundamental components of detection systems, and acceleration techniques in object detection. The author emphasized the substantial impact that deep learning has had on object detection and its role in driving the technological advancement of this field.

Sreenu and Durai [9] conducted an extensive survey of deep learning techniques applied to crowd analysis in video surveillance. The authors focused on reviewing existing research in the field, covering topics such as action recognition, object recognition, violence detection, and crowd analysis in crowded environments. Their survey involved a comparison of various deep learning algorithms and models, examining their potential applications in real-time crowd video processing. Also identified the current limitations and challenges of existing methods and suggested future research directions to address these obstacles. The authors conducted the survey by analyzing a bibliographic summary of papers from various digital libraries, including ScienceDirect, IEEE Xplore, and ACM Digital Library.

Wu *et al.* [10] undertaken a detailed analysis of recent advances in deep learning-based object detection. The authors covered a range of factors that influence detection performance, and discussed potential research directions in the field. The authors highlighted the significance of deep convolutional neural networks as a critical technology for enhancing object detection performance.

Liu *et al.* [11] provided an overview of recent advances in the domin of generic object detection using deep learning methods. More than 300 research papers on a variety of topics related to object detection, including detection frameworks, feature representation, proposal generation, context modelling, training methodologies, and evaluation metrics, were assessed by the authors. The authors also suggested intriguing lines of inquiry for additional study in the area.

Wang *et al.* [12] performed a detailed survey of recent developments in Salient Object Detection (SOD) in the deep learning era. The authors conducted a thorough analysis of deep SOD (Salient Object Detection) algorithms, considering various aspects and also evaluated existing SOD datasets and assessment metrics. The authors also performed a comprehensive benchmarking of a wide range of representative SOD models. Additionally, they examined the performance of SOD algorithms under different attribute settings, the resilience of SOD models to adversarial attacks and erratic input perturbations, as well as the generalization and complexity of existing SOD datasets. The paper concluded by discussing outstanding issues and potential research directions in the field of SOD.

Zhao et al. [13] conducted an extensive study of deep learning-based frameworks to detect object, covering a wide range of topics. The authors provided a historical account of deep learning and the use of CNN in this field. They also presented an overview of common architectures for object detection, along with various techniques and modifications to improve detection performance. Additionally, the paper included a brief survey of specific tasks, such as face detection, pedestrian detection, and salient object detection. The authors performed experimental analyses to compare different methods, and concluded with a discussion of potential directions and future research tasks for object detection and other neural network-based learning systems.

4. RESEARCH OBJECTIVES

1. **To critically evaluate** the performance of state-of-the-art deep learning-based object detection models in video surveillance.
2. **To identify** the strengths and weaknesses of existing models, including YOLO, Faster R-CNN, RetinaNet, and SSD, in handling

occlusions, real-time constraints, and varying environmental conditions.

3. **To propose and develop** an optimized deep learning-based object detection framework that balances speed, accuracy, and computational efficiency.
4. **To explore the integration** of multiple detection methodologies to enhance detection robustness, particularly in complex surveillance environments.
5. **To assess the feasibility** of deploying object detection algorithms in real-world scenarios such as smart city surveillance while addressing privacy and security concerns.

5. RESEARCH QUESTIONS

1. What are the key limitations of existing deep learning-based object detection models in video surveillance?
2. How do different object detection architectures compare in terms of accuracy, speed, and adaptability to real-world scenarios?
3. What are the most effective strategies to enhance the real-time performance and robustness of object detection models?
4. How can multiple detection methodologies be integrated to improve detection efficiency in surveillance applications?
5. What are the practical challenges and solutions for implementing deep learning-based object detection in real-world surveillance systems?
6. What are the most recent deep learning approaches and their performance status in object detection for video surveillance?
7. What are the recent advancements and technical improvements in multi-object tracking for video surveillance?
8. What are the recent progress and innovations in scene understanding and activity recognition in video surveillance through deep learning?
9. What are the ongoing technical challenges and limitations in the area of object detection and object tracking for video surveillance and how can they be addressed?

6. REVIEW METHODOLOGY

In this paper, we followed a systematic literature review methodology to investigate a research question and achieve the review's objectives. The methodology involved defining the research question and objectives, which helped in defining the scope of the review and criteria for inclusion and exclusion of studies.

- We conducted an extensive search for relevant literature using various databases and sources, guided by well-defined keywords and inclusion and exclusion criteria.
- After the initial search, we screened the studies based on their titles and abstracts and selected the relevant studies that met the inclusion criteria and were pertinent to the research question and objectives.
- Pertinent data was extracted from the selected studies and subsequently synthesized through a qualitative or quantitative synthesis.
- The results were presented using various methods, such as a narrative report, tables, and figures. In conclusion, the main findings of the review were summarized, and suggestions for future studies were provided.

The PRISMA reporting standard was followed to outline a checklist of items to be reported in the systematic review, including search methods, inclusion/exclusion criteria, and data extraction/synthesis methods.

6.1 Research Design

This study employs an **experimental research design** to analyze and compare various deep learning models for object detection. The models are evaluated based on accuracy, precision, recall, and computational efficiency using publicly available and custom-built datasets.

6.2 Data Collection and Preprocessing

Data Sources

- Public datasets: COCO, Pascal VOC, Cityscapes, KITTI, SOC, IR Thermal Image datasets, etc.
- Real-world data collection: If applicable, surveillance videos/images collected from specific environments (e.g., drone footage, security camera feeds).

Data Preprocessing Steps

- **Annotation:** Bounding boxes or segmentation masks labeled using LabelImg or VIA (VGG Image Annotator).
- **Normalization:** Pixel values scaled between 0 and 1.

- **Augmentation:** Random rotations, flipping, contrast adjustments to enhance model generalization.
- **Splitting:** Train-Test-Validation split (typically 70%-20%-10%).

6.3 Model Selection and Training Deep Learning Models Evaluated

- CNN-based architectures (ResNet, YOLO, Faster R-CNN, RetinaNet, SqueezeNet).
- Hybrid architectures (OF-ConvAE-LSTM, DaCoLT).
- Experiment with different backbone networks (e.g., ResNet-50, VGG16, MobileNet).

6.4 Model Evaluation Metrics

Models are evaluated using the following metrics:

- **Accuracy:** Correctly detected objects vs. total objects.
- **Precision & Recall:** How well the model identifies true positives and minimizes false positives.
- **Mean Average Precision (mAP):** Area under the precision-recall curve.
- **Inference Speed:** Frames per second (FPS) for real-time detection.

6.5 Validation and Testing

- Cross-validation (5-fold validation) for robustness.
- Comparison of models across multiple datasets for generalization.
- Real-world testing on live video feeds (if applicable).

7. ARTICLE SELECTION STRATEGY (ASS)

Article selection is a crucial step in Systematic Literature Review (SLR) as it determines the scope and quality of the review. To conduct a review, the selection process of determining which articles to include or exclude requires establishing specific criteria. The inclusion criteria include the scope of the topic, publication date, language, and type of publication. The exclusion criteria may include articles that are not relevant to the research question, articles without a deep learning focus, and articles that are duplicates of others. The articles are then screened, and the relevant articles are selected for further analysis and synthesis. The purpose of this process is to ensure that the review includes only the most pertinent and current articles, thereby offering a comprehensive and precise reflection of the current state of the field.

8. DATA SYNTHESIS AND ANALYSIS

DEEP LEARNING MODELS

Video surveillance object detection is a significant and dynamic research area, with numerous algorithms under development and experimentation. Among the prevalent techniques in object detection for video surveillance systems, deep learning models are extensively employed. Some of the commonly adopted deep learning models for object detection include:

YOLO

In recent years, object detection system YOLO (You Only Look Once) has become increasingly popular as a real-time solution in video surveillance applications for object detection. YOLO has been described by He *et al.* [14] as a single-shot detection (SSD) algorithm that analyzes full image in a single forward pass, resulting in faster and more efficient performance compared to other object detection algorithms. This speed advantage has contributed to YOLO's widespread use in video surveillance applications, where quick object detection is crucial for effective monitoring and security [15].

However, despite its popularity, YOLO is not without its limitations. One of the major criticisms of YOLO is its computational intensity, which makes its inference speed slower than other object detection algorithms. This can be an issue for real-time applications, such as video surveillance, where quick processing is necessary. Additionally, YOLO struggles with detecting objects in low-resolution images and is prone to false positive detections due to background clutter or overlapping objects. The algorithm is also limited in terms of customizability, which means that users cannot easily modify it to suit their specific needs or preferences [16].

While YOLO has gained widespread popularity due to its real-time performance and high accuracy, it is important to consider its limitations when evaluating its suitability for specific applications. The fast processing speed and end-to-end design of YOLO make it a valuable tool for object detection in video surveillance, but its computational intensity and limited customizability must also be taken into account.

CNN

A widely adopted deep learning architecture for object detection in video surveillance is the Convolutional Neural Network (CNN) [17]. Due to their capacity to learn intricate features from input data, CNNs have demonstrated high effectiveness in object detection tasks. They are capable of handling large-scale image datasets, making them well-suited for video surveillance applications where large amounts of data are generated and also it can be an end-to-end training, allowing them to learn both the feature extraction and classification tasks automatically. This can be used for detecting object in real-time video streams, allowing for real-time monitoring and alerting in surveillance applications. They can also be trained on a large number of classes, making them useful for multi-object detection tasks.

With the availability of large pre-trained models, such as VGG and ResNet, CNNs can also be used for transfer learning, which can save time and computational resources. The main motivation behind using this in object detection is their ability to identify high-level features from raw data, such as images, and make predictions based on these features. The CNN is capable of handling and processing large volumes of data, and is able to learn from it. This is crucial for detecting object in video surveillance, where the number of frames can be substantial. CNNs can learn from these frames and extract meaningful features from them, which can be used to identify objects in the video [18].

In addition, CNNs have been shown to be robust and effective in object detection, even when dealing with complex and cluttered scenes. They can also handle scale and rotation variations, which is important in video surveillance, where the scale and orientation of objects can change significantly over time [19]. Despite these advantages, there are also some limitations to using CNNs for object detection in video surveillance [20]. Despite their effectiveness, CNNs have some notable limitations. For instance, one of the primary requirements for training the network is large amounts of labeled data, which can be a difficult and more time-consuming task to acquire. Furthermore, CNNs can be computationally expensive, posing a significant challenge in video surveillance for detecting real-time object. In conclusion, CNNs are a powerful tool for detecting object in video surveillance, but careful consideration must be given to their limitations and requirements.

R-CNN

The Recurrent Convolutional Neural Network (RCNN) is a deep learning algorithm used for object detection in computer vision applications [21]. It combines the strengths of CNNs and RNNs to effectively handle complex, multi-object scenes and achieve improved accuracy over traditional object detection methods [22]. RCNNs have been widely adopted in video surveillance, self-driving vehicles, and security systems due to their ability to perform accurate object detection even in cluttered or complex scenes [23]

However, RCNNs also have a few limitations, including high computational cost, longer training times, and sensitivity to the quality of region proposals [24]. Additionally, the large number of parameters in RCNNs can also make them more prone to overfitting, negatively impacting performance on new data. Despite these limitations, RCNNs are expected to continue to play a significant role in object detection and related computer vision tasks, with researchers constantly exploring ways to improve their performance, such as increasing processing speed and accuracy [23]. The advancement of hardware and software technologies is expected to lead to wider adoption of RCNNs in various applications, including video surveillance [21].

SDD

Single Shot MultiBox Detector (SSD) is an object detection algorithm based on deep learning that is widely used in computer vision applications [25]. It was first introduced by Liu *et al.* [26]. The main advantage of SSD is its speed and real-time performance, which make it an ideal choice for use in real-time applications such as video surveillance, self-driving cars, and others. SSD is designed to detect objects of various sizes and shapes in a single shot, rather than applying object detection multiple times like in the case of RCNN-based algorithms. This reduces the computational cost and makes SSD faster compared to other object detection algorithms.

In real-world applications, the need for SSD arises due to the increasing demand for speed and accurate object detection algorithms. As video surveillance and self-driving cars require real-time object detection, SSD has gained popularity in these fields [27]. SSD's strength lies in its real-time processing capabilities, which makes it well-suited for deployment in a kind of practical applications such as self-driving cars and video surveillance [28]. SSD's high speed and accuracy compared to other object detection algorithms has made it an attractive

option for many computer vision use cases. Due to the increasing demand for fast and accurate object detection algorithms, SSD's future growth is expected to be significant, particularly in domains like video surveillance and self-driving cars. Researchers are continuously working to refine SSD's accuracy and performance, and it is anticipated that novel techniques and alterations to the algorithm will emerge in the near future [29].

DNN

Deep Neural Networks (DNNs) are machine learning models that use artificial neural networks with multiple hidden layers. They have gained widespread popularity for computer vision and image processing tasks, such as object detection, classification, and recognition. DNNs aim to automatically extract high-level features from raw data, allowing these features to be robust to variations in the data [30].

DNNs can be trained with extensive data to learn intricate connections between input features and output labels, and fine-tuned using transfer learning to enhance their performance for specific tasks. In the area of object detection, DNNs have been employed in different forms, including Fast R-CNN, Faster R-CNN, and RetinaNet, to achieve state-of-the-art outcomes on standardized datasets [31].

DNNs are particularly valuable for object detection tasks due to their ability to learn intricate relationships between input features and output labels, as well as their effectiveness in processing large amounts of data. DNNs can be trained end-to-end, enabling the learning process to be optimized from input to output layers, which can reduce the need for feature engineering efforts.

DNNs have broad applications for object detection across various domains, including images, videos, and real-time scenarios. DNNs are often used in conjunction with other techniques, such as reinforcement learning, to enhance their object detection performance.

The future growth of DNN in object detection is expected to continue, as researchers continue to explore new techniques to improve the accuracy, efficiency, and scalability of DNNs [32]. This could include the use of new architectures, such as

transformers, as well as the development of new algorithms and approaches to handle large amounts of data. The increasing availability of efficient performance in computing resources, includes GPUs and TPUs, will also enable the training of larger and more complex DNNs [33].

AR-NET

The AR-Net (Appearance and Motion Features Regression Network) algorithm is a deep learning-based object detection approach specifically developed for video-based object detection tasks. It combines both appearance and motion information to overcome motion-related challenges [34]. AR-Net uses a CNN to extract appearance features from a single frame and a LSTM to analyze temporal information from a sequence of frames, leading to superior performance in object detection compared to methods that solely rely on appearance features [35].

The scope of AR-Net lies in object detection application in video surveillance, self-driving cars, and other video-related tasks. Its use in video object detection is particularly useful in scenarios where objects are moving and traditional object detection algorithms based on appearance only may fail to accurately detect them [36].

AR-Net has limitations that include its dependency on the quality of the given input data. Poor quality or noisy video frames can negatively impact the algorithm's performance. Additionally, the algorithm may not function effectively when objects are occluded or undergo significant changes in appearance over time.

AR-Net is a promising algorithm for video object detection that has the potential for future growth and application in various video-related tasks [37]. Further research and development are needed to overcome its limitations and improve its performance in real-world applications.

Table 1: A Detailed comparison showing PMI for various models.

Model	Plus (Strengths)	Minus (Weaknesses)	Best Model for (Interesting Fact)
YOLO (v1-v4, Fast YOLO)	• Fast inference speed • High accuracy in real-time tasks (e.g., 97.93% AP in thermal images, 91.73% mAP for weapon detection)	• Struggles with small objects in cluttered scenes • Limited performance on complex datasets	Real-time object detection (e.g., surveillance, vehicle tracking)

Faster R-CNN	• High precision in complex environments (up to 97.23% accuracy in surveillance tasks)	• Computationally intensive; slower than YOLO	Security applications requiring precision (e.g., firearm detection)
RetinaNet	• Strong precision in security tasks (~97%) • Effective in detecting small objects	• Requires fine-tuning for optimal speed and accuracy	Security surveillance, object detection in challenging conditions
ResNet (e.g., ResNet-101)	• Strong generalization for various domains • Effective in drone detection (though accuracy is lower at 40.99%)	• Lower accuracy in drone detection compared to specialized models	General object recognition; drone detection (with optimization)
VGG16, SqueezeNet	• Strong accuracy (93.8%) in urban multi-object detection • Well-suited for classification tasks	• May require additional layers for complex object detection	Multi-object detection in urban environments
OF-ConvAE-LSTM	• Excellent in anomaly detection for video surveillance (96.5% accuracy)	• Requires extensive training data • May struggle in non-sequential data tasks	Anomaly detection in crowd surveillance and security footage
Modified YOLOv1	• Efficient in simpler object detection tasks	• Poor performance on complex datasets (65.6% accuracy on Pascal VOC)	Basic object detection in low-data scenarios
DaCoLT	• Enhances CNN performance using contrast and darkening techniques	• May require additional model integration for optimal results	Low-light object detection; enhanced image clarity in security settings

Table 2: Object Detection Using Deep Learning

S. No.	Author	Proposed Deep learning Model	Type of Detection	Data Set Used	Accuracy's
1	Duman & Erdem (2021) [38]	OF-ConvAE-LSTM	Anomaly Detection (crowded scenes) in Videos	Avenue, UCSD Ped1, UCSD Peds2	89%, 96.5% respectively
2	Salido et al., (2021) [39]	RetinaNet & YOLOv3	In Video Surveillance Images, Automatic Handgun Detection	MS COCO	RetinaNet: Average Precision and Recall 96.36%, 97.23% YOLOv3: Precision and F1 Score 96.23%, 93.36%
3	Castillo et al. (2018) [40]	CNN & Darkening and Contrast at Learning and Test stages (DaCoLT)	Automatic Cold Steel Weapon Detection	COCO	87.74%
4	Nalamati et al., (2019) [41]	ResNet-101 and Inception with Faster-RCNN, SSD	Drone Detection in Surveillance Videos	COCO	40.99%
5	Kwan et al., (2020) [42]	YOLO	Vehicle Detection and Classification	YOLO	76%
6	Solovyev, Wang & Gabruseva 2021 [43]	RetinaNET – Resnet	Object detection	Microsoft COCO, Open Images	56.1 mAP - validation data set, 56.4 mAP - test-dev set
7	Boudjit & Ramzan (2021) [44]	YOLO-v2	Person Detection		98% (object detection) and 96.5% (target object tracking)

8	Bhatti et al., 2021 [45]	Yolov4	Weapon Detection	Constructed training database for real-time scenario	91.73% mean average precision (mAP), F1-score of 91% with almost 99% confidence score
9	Rashid et al., 2020 [46]	CNN	Object Recognition	Caltech-101, Butterflies database, Birds database, and CIFAR-100	95.5%, 100%, 98%, and 68.80%
10	Krišto et al., 2020 [47]	YOLOv3	Object Detection in Thermal Images	IR thermal image dataset and so on	97.93% Average Precision (AP) for the Person class in all weather conditions
11	Lu et al., 2019 [48]	Fast YOLO	Object Detection in Videos	Vehicle Monitoring Dataset obtained from the Xiamen municipal transportation bureau	88.45%
12	Fu et al., 2019 [49]	CNN / Deepside	Salient Object Detection	SOC dataset	88.50%
13	Ahmad et al. (2020) [50]	Modified YOLOv1 Neural Network	Object Detection	Pascal VOC datasets 2007/2012	65.6% and 58.7%
14	Li et al., (2021) [51]	Improved Faster R-CNN (VGG16-based)	Multi-Object Detection and Recognition	Cityscapes, KITTI	93.8%
15	Shorfuzzaman et al., 2021 [52]	Faster R-CNN, SSD, YOLO	Social distancing violation	real-world datasets	84.29% - 86.87%
16	Yang et al., 2019 [53]	Faster R-CNN with Spatio-Temporal Information (SLD)	In THz Security Images, suspicious Object (bottle, gun, knife) Detection and Recognition	COCO	85.65% - 97.82%
17	Chin et al., (2019) [54]	MXNet	Video Object Detection	ImageNet VID and mini YouTube-BoundingBoxes	75.5%
18	Muhammad et al., (2018) [55]	CNN based on SqueezeNet Architecture	Fire detection and localization	Two benchmark datasets	92.59%

Table:3 Object Tracking comparison

S. No	Author	Proposed Deep learning Model	Data Set Used	Accuracy's
1	Chen et al., (2019) [56]	YOLO V3 and SSD	Cityscape, KITTI	85%
2	Goyal et al., 2020 [57]	Viola-Jones algorithm	Collection of positive and negative images	Average 94.93% for human detection, 95.2% for vehicle detection, 97.67% for weapon detection, 97% for overall accuracy (compared to 64% for previous method)
3	Zhang (2017) [58]	Recurrent Convolutional Neural Network (RCNN) with Reinforcement Learning (RL)	PASCAL VOC dataset & ageNet dataset	63.5%

4	Zhang et al., 2021 [59]	FairMOT	six training datasets - ETH, CityPerson The CalTech, MOT17 , CUHK-SYSU and PRW	79.4%
5	Papakis et al., 2021 [60]	Graph Convolutional Neural Network (GCNN)	MOT challenge dataset	64.5%
6	CAO et al., 2018 [61]	Knowledge-Guided Training Framework for Deep Neural Networks Vehicle Object Detection Support Vector Machine (SVM).	CIFAR-10, ImageNet	66.9%
7	Gan et al., 2018 [62]	CNN-based Multiple Object Tracking	MOT17 and MOT16 benchmark datasets	78.9%
8	Zhang et al., 2017 [63]	CNN with Spatial-Temporal Saliency Guided Sampling	A set of datasets for tracking objects that are difficult to follow due to their non-rigid or generic nature.	77%
9	Hu et al., 2020 [64]	CRAM (Convolutional Regression Network with Appearance and Motion Features)	ImageNET experimental satellite video dataset	72.86%
10	Elhoseny, 2020 [65]	MODT (Multi-object Detection and Tracking)	Kalman Filtering	76.23% (Detection), 86.78% (Tracking)

Table 4: Activity Recognition Or Scene Understanding

Author	**Proposed Deep learning Model**	**Type of Activity Recognition**	**Data Set Used**	**Accuracy's**
Zhang et al., 2020 [66]	Simplified Temporal Convolutional Network (STCN)	Construction Equipment Action Recognition (CEAR)	CEAR	76.25%
Ullah et al., 2018 [67]	Optimized Deep Autoencoder (DAE) & CNN	Action recognition	UCF50, UCF101, HMDB51, YouTube Action dataset	92.3%, 89.7%, 96.4%
Dua et al. 2020 [68]	Multi-input CNN-GRU	Human Activity Recognition	UCI-HAR, WISDM, PAMAP2	96.20%, 97.21%, 95.27%
Jaouedi et al., 2019 [69]	Hybrid Deep Learning Model	Human Action Recognition	UCF101, UCF Sports, KTH	96.3%
Khan et al., 2020 [70]	Combination of deep neural network (DNN) and multiview features	Human Action Recognition (HAR)	HMDB51, YouTube, IXMAS, UCF Sports, KTH	93.7%, 98%, 99.4%, 95.2%, 97%
Wan et al., 2019 [71]	CNN	Human Activity Recognition	UCI and Pamap2	82% - 92%
Mliki et al., 2020 [72]	Convolutional Neural Network (CNN)	Human Activity Recognition	UCF-ARG	99.5%

Xia et al., 2020 [73]	Long Short-Term Memory (LSTM)-CNN	Human Activity Recognition	UCI-HAR, WISDM, OPPORTUNITY	UCI-HAR -95.78%, WISDM -95.85%, OPPORTUNITY - 92.63%
Xing et al., 2019 [74]	Deep Convolutional Neural Networks (CNN)	Driver Activity Recognition	AlexaNet, GoogleNet, ResNet	AlexNet: 81.6%, GoogLeNet: 78.6%, ResNet50: 74.9% (for 7 tasks), Binary Classification (Distracted/Not Distracted): 91.4%
Meng et al., 2020 [75]	AR-Net	Action Recognition	ActivityNet v1.3, FCVID, Mini-Kinetics	79.7%

8. DISCUSSION

We have reviewed in-depth a variety of deep learning methods for object detection in various scenarios, including video surveillance, real-time applications, industrial systems, and environmental conditions such as traffic, weather, and fire. The reviewed papers presented various approaches, including Convolutional Autoencoder (CAE), Deep Learning (DL), YOLO, Faster R-CNN, and other methods, to detect and recognize objects such as handguns, cold steel weapons, drones, vehicles, humans, salient objects, and others. The literature also explored techniques to improve object detection performance, such as brightness-guided preprocessing, content-aware guidance, cross-layer fusion, and others. The papers aimed to contribute to the development of sustainable smart cities through mass video surveillance and real-time object detection for security, safety, and industrial applications.

9. METRICS USED

Here is the comparison of the various metrics you mentioned for different models of deep learning used for object tracking, object detection, and scene understanding in video surveillance.

Table: 5 Metrics Used for Measuring the Performance Analysis

S.No.	Metrics	Description
1.	F-score	F-score quantifies a model's accuracy and precision, taking into account both the false positive and false negative rate.
2.	Precision	The measure of precision determines the ratio of accurate positive identifications to all the positive identifications anticipated by the model.
3.	False Positive (FP)	A false positive is a prediction made by the model that is not actually an object of interest.
4.	Intersection over Union (IoU)	The level of overlap between the bounding box predicted by the model and the actual or ground truth bounding box is evaluated using IoU metric.
5.	Mean Average Precision (mAP)	The average precision across all object classes in the dataset is measured by mAP.
6.	Recall	The proportion of accurate positive identifications out of all the objects present in the dataset is evaluated by Recall.
7.	Area Under the ROC Curve (AUC)	The capability of a model to differentiate between positive and negative examples is gauged by AUC
8.	Computational Complexity	To run the model, it is important to consider the quantity of computational resources and time required.

10. PERFORMANCE OF DL TECHNIQUES OBJECT DETECTION:

Based on the given performance scores, it can be concluded that YOLOv3 and YOLO-v2 are the top-performing algorithms in object detection, followed by RetinaNet and OF-ConvAE-LSTM. Convolutional Neural Networks (CNNs) also show good performance, with scores ranging from 87.74% to 99%. On the other hand, R-CNNs have lower performance, with scores ranging from 35.00% to 93.80%. SSD has the lowest performance among the

algorithms, with scores of 15.00% and 44.50%. It's important to note that the performance of these algorithms can differ based on the specific problem and the types of objects being detected.

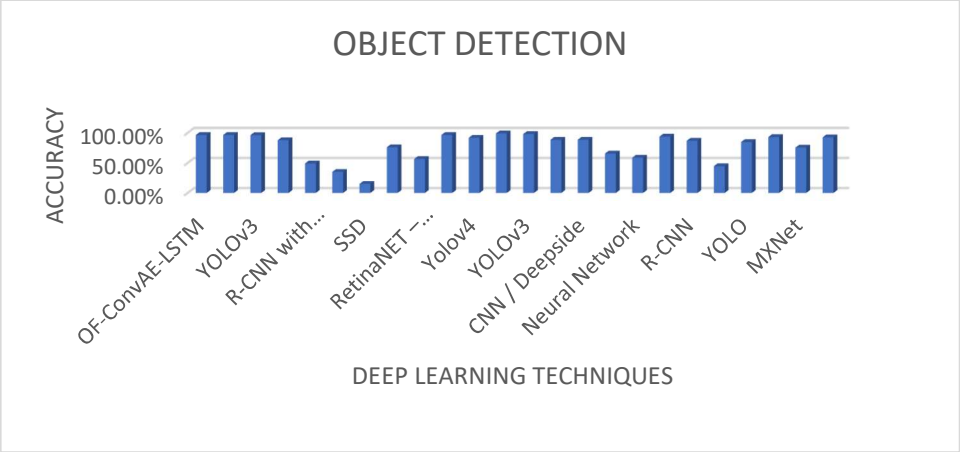

Figure 2: Performance Analysis Of Various DL Techniques Used In Object Detection

The Viola-Jones algorithm has the highest accuracy among the listed object detection algorithms, with a score of 97%. YOLO V3 and MODT also perform well, with scores of 85% and 86.78% respectively. SSD, FairMOT, and CRAM have similar performance scores, ranging from 79.50% to 72.86%. RCNN and GCNN have lower performance scores of 63.50% and 64.50%. SVM has the lowest performance, with a score of 66.90%. CNNs have performance scores of 78.90% and 77%. These scores can differ depending on the specific problem and type of objects being identified.

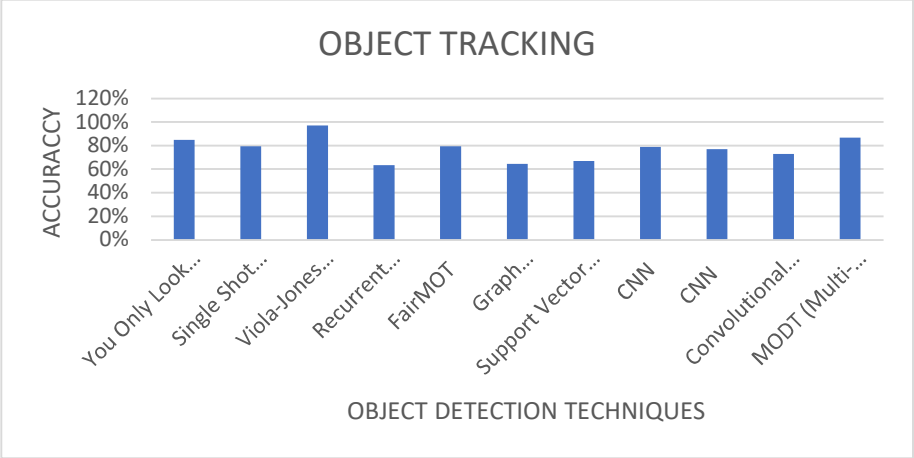

Figure 3: Performance Analysis Of Various DL Techniques Used In Object Tracking

ACTIVITY RECOGNATION:

The Deep Neural Network (DNN) model has the highest accuracy among the listed algorithms, with a score of 98%. The CNN-GRU and Hybrid Deep Learning Model also perform well, with scores of 97.21% and 96.30% respectively. The DAE & CNN and LSTM-CNN models have similar performance scores of 96.40% and 95.85%. The first CNN model has an accuracy of 92%, while the second CNN model has a significantly higher accuracy of 99.50%. The CNN model has an accuracy of 81.60%, and the AR-Net model has a lower accuracy of 79.70%. The STCN model has an accuracy of 76.25%. These scores can vary depending on the specific problem and type of data being analyzed.

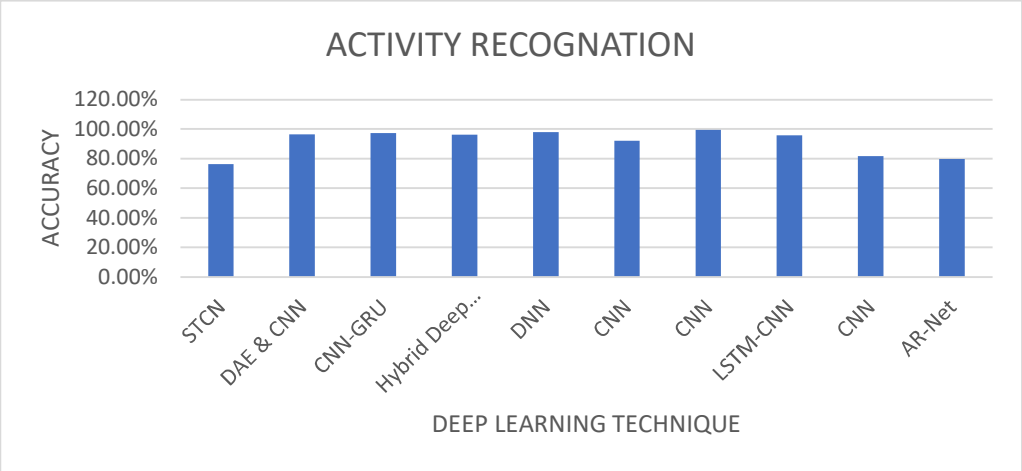

Figure 4: Performance Analysis Of Various DL Techniques Used In Scene Understanding

In summary, the success of algorithms designed for detecting objects, tracking them, and recognizing actions can be influenced by the nature of the data being analyzed, as well as the specific problem being addressed. For object detection, YOLOv3 and YOLO-v2 perform best, followed by RetinaNet and OF-ConvAE-LSTM. For object tracking, the Viola-Jones algorithm has the highest accuracy, followed by YOLO V3 and MODT. For action recognition, the DNN model performs best, followed by the CNN-GRU and Hybrid Deep Learning Model. Overall, it can be seen that Deep Neural Network models and Convolutional Neural Networks tend to perform well in these tasks, with performance scores ranging from 78.90% to 98%.

11.RESEARCH IMPLICATION AND PRACTICE:

The research papers you have reviewed generally focus on object detection in various applications, including video surveillance, UAVs, smart manufacturing, and traffic environment using deep learning techniques. The studies utilize algorithms such as optical flow, convolutional autoencoders, YOLO, and improved Faster R-CNN to enhace object detection speed and accuracy. The research implications of these studies suggest that deep learning techniques can effectively be used for object detection tasks in various domains. The results of these studies have practical implications in field includes security and surveillance, traffic management, and industrial automation.

The papers we reviewed were primarily focused on object detection in video surveillance images and videos, with an emphasis on deep learning methods. Several of the research papers focus on identifying weapons, vehicles, and humans in video surveillance footage. Many of these studies employ deep learning models to improve detection accuracy and speed. Optical flow and attention mechanisms are also used to enhance algorithm performance, indicating potential areas for further investigation. The overall implication of this research is that there is a rising trend in using deep learning for object detection in video surveillance, which holds promise for achieving better results.. There is a need for further research to extend object detection algorithms to challenging scenarios, such as in adverse weather conditions or in the presence of small objects and the practices are Object detection algorithms can be used in video surveillance systems to provide real-time monitoring and analysis, which can help improve security and prevent crime. Combining deep learning algorithms with technologies such as THz imaging can improve security in video surveillance. Attention mechanisms can also enhance the robustness of object detection algorithms and facilitate more effective analysis of video surveillance data.

In conclusion, further research is required to enhance the speed and accuracy of object detection systems. The use of deep learning techniques in object detection offers significant potential to advance these fields.

12. RESEARCH HYPOTHESIS

"Advanced deep learning models, particularly those utilizing CNN-based architectures (YOLO, RetinaNet, Faster R-CNN), demonstrate superior accuracy in object detection tasks across diverse datasets, with variations in performance attributed to dataset complexity and model optimizations."

13. CHALLENGE AND STRENGTH:

Strengths of detecting object in video surveillance using Deep Learning are mmproved accuracy using Deep learning algorithms can accurately detect objects in complex scenes, with fewer false positives and false negatives compared to traditional computer vision techniques and the real-time performance using deep learning algorithms can process video frames in real-time, making them suitable for real-world video surveillance applications and Handling variations in object appearance in deep learning algorithms can handle significant variations in object appearance, such as changes in illumination, scale, and viewpoint, making them suitable for challenging video surveillance scenarios.

Object Detection using deep learning challenges in Video Surveillance are Computational complexity in Deep learning algorithms can be computationally intensive, requiring high processing power and large amounts of memory and large quantity of training data for deep learning algorithms to learn, enormous amounts of annotated training data are needed to detect objects accurately, which can be time-consuming and expensive to obtain and the limited generalization using deep learning algorithms may not generalize well to new scenarios and objects that were not included in the training data. Deep learning algorithms have limited interpretability, making it hard to comprehend their decision-making process, potential errors, and biases. The presence of bias and fairness issues in deep learning algorithms is also a concern, as they may reflect biases from the training data, resulting in discriminatory or unjust outcomes.

13. LIMITATIONS:

Some limitations of exploring the advancements and challenges of object detection in video surveillance through deep learning include limited availability of large-scale annotated datasets for computational resources required for training deep learning models can be expensive and demanding and also high false positive and false negative rates in object detection, which can limit the practical applications of these models and further limited generalizability of deep learning models to new environments and scenarios, due to their reliance on training data that may not accurately represent real-world conditions and also privacy concerns with using video surveillance, especially when it comes to collecting and using personal data. Even in technical challenges in integrating object detection with existing video surveillance systems and making the results actionable. Also, the ethical concerns with the use of object detection in video surveillance, including bias, discrimination, and the potential for misuse of the technology and the ongoing advances in object detection and video surveillance technology can quickly make current models and systems obsolete, requiring constant updating and maintenance.

14. CONCLUSION AND FUTURE RESEARCH DIRECTIONS:

In this paper we presented a detailed survey covering many object detection models. We compared different types of object detection algorithms in video surveillance with their merits and demerits such as YOLO, CNN, R-CNN, Faster R-CNN, Retina Net, SSD etc. This review paper provides a detailed analysis of state-of-the-art object detection models, architectures and evaluates the performance of models using standard datasets. We determined Deep learning-based object detection has been a research area in recent years. It is the rapidly growing field with great potential for real-world applications. Despite the advancements in this field, there are still many challenges that need to be addressed, such as real-time processing, accuracy, handling large-scale datasets, and incorporating domain-specific knowledge. There are significant prospects for further investigation and improvement in these fields. It can also be deduced that the use of deep learning methods for detecting objects in video surveillance has led to considerable enhancements in the accuracy and speed of detection. However, there are still challenges that need to be addressed such as dealing with occlusions, enhancing real-time efficiency, and increasing the scope of object classes that can be detected. Future research directions in this field can include further improvement for object detection deep learning algorithms, exploring the integration of multiple object detection methods, and the development of more robust and efficient real-time detection systems. Additionally, research could also focus on the deployment of object detection in practical scenarios such as in smart cities and improving the privacy and security concerns associated with the use of video surveillance. Object detection algorithms can be made more robust by incorporating domain-specific knowledge such as scene context, object properties, and prior knowledge. Future research can focus on developing methods that incorporate such knowledge to enhance object detection algorithms performance in video surveillance.

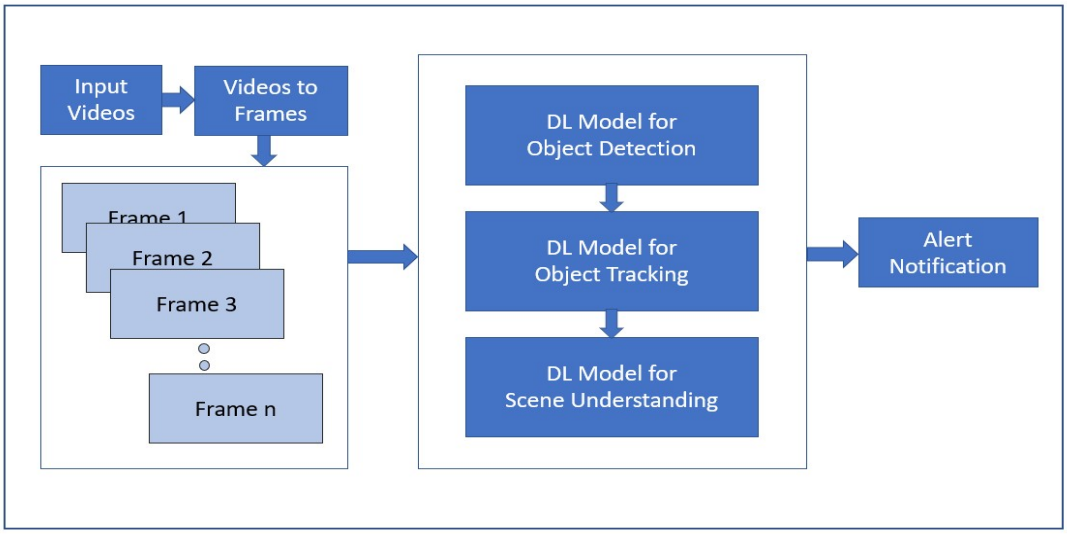

Figure 5: Future Research Directions For Video Surveillance Using Deep Learning

REFERENCE

[1] T. D. Räty, "Survey on contemporary remote surveillance systems for public safety," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, no. 5, pp. 493-515, 2010.

[2] I. R. I. Haque and J. Neubert, "Deep learning approaches to biomedical image segmentation," Informatics in Medicine Unlocked, vol. 18, 100297, 2020.

[3] J. Luo, Z. Yang, S. Li, and Y. Wu, "FPCB surface defect detection: A decoupled two-stage object detection framework," IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-11, 2021.

[4] M. Längkvist, L. Karlsson, and A. Loutfi, "A review of unsupervised feature learning and deep learning for time-series modeling," Pattern recognition letters, vol. 42, pp. 11-24, 2014.

[5] S. Salagrama, H. H. Kumar, R. Nikitha, G. Prasanna, K. Sharma, and S. Awasthi, "Real time social distance detection using Deep Learning," in 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), pp. 541-544, May 2022, IEEE.

[6] S. Saponara, A. Elhanashi, and A. Gagliardi, "Implementing a real-time, AI-based, people detection and social distancing measuring system for Covid-19," Journal of Real-Time Image Processing, pp. 1-11, 2021.

[7] N. Ghatwary, M. Zolgharni, and X. Ye, "Early esophageal adenocarcinoma detection using deep learning methods," International journal of computer assisted radiology and surgery, vol. 14, pp. 611-621, 2019.

[8] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," Proceedings of the IEEE, 2023.

[9] G. Sreenu and S. Durai, "Intelligent video surveillance: a review through deep learning techniques for crowd analysis," Journal of Big Data, vol. 6, no. 1, pp. 1-27, 2019.

[10] X. Wu, D. Sahoo, and S. C. Hoi, "Recent advances in deep learning for object detection," Neurocomputing, vol. 396, pp. 39-64, 2020.

[11] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," International journal of computer vision, vol. 128, pp. 261-318, 2020.

[12] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 6, pp. 3239-3259, 2021.

[13] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, "Object detection with deep learning: A review," IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 11, pp. 3212-3232, 2019.

[14] W. He, Z. Huang, Z. Wei, C. Li, and B. Guo, "TF-YOLO: An improved incremental network for real-time object detection," Applied Sciences, vol. 9, no. 16, pp. 3225, 2019.

[15] E. C. Joseph, O. Bamisile, N. Ugochi, Q. Zhen, N. Ilakoze, and C. Ijeoma, "Systematic Advancement of Yolo Object Detector For Real-Time Detection of Objects," in 2021 18th International Computer Conference on Wavelet

Active Media Technology and Information Processing (ICCWAMTIP), December 2021, pp. 279-284, IEEE.
[16] S. Du, B. Zhang, P. Zhang, P. Xiang, and H. Xue, "FA-YOLO: an improved YOLO model for infrared occlusion object detection under confusing background," Wireless Communications and Mobile Computing, 2021, pp. 1-10.
[17] A. R. Pathak, M. Pandey, and S. Rautaray, "Application of deep learning for object detection," Procedia Computer Science, vol. 132, pp. 1706-1717, 2018.
[18] D. Han, Q. Liu, and W. Fan, "A new image classification method using CNN transfer learning and web data augmentation," Expert Systems with Applications, vol. 95, pp. 43-56, 2018.
[19] Y. Ji, H. Zhang, Z. Zhang, and M. Liu, "CNN-based encoder-decoder networks for salient object detection: A comprehensive review and recent advances," Information Sciences, vol. 546, pp. 835-857, 2021.
[20] D. T. Nguyen, T. N. Nguyen, H. Kim, and H. J. Lee, "A high-throughput and power-efficient FPGA implementation of YOLO CNN for object detection," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 27, no. 8, pp. 1861-1873, 2019.
[21] Z. He and L. Zhang, "Domain adaptive object detection via asymmetric tri-way faster-rcnn," in Proceedings of the 16th European Conference on Computer Vision, Glasgow, UK, 2020, Part XXIV, pp. 309-324.
[22] H. Zhang, H. Chang, B. Ma, N. Wang, and X. Chen, "Dynamic R-CNN: Towards high quality object detection via dynamic training," in Proceedings of the 16th European Conference on Computer Vision, Glasgow, UK, 2020, Part XV, pp. 260-275.
[23] V. Mazzia, A. Khaliq, and M. Chiaberge, "Improvement in land cover and crop classification based on temporal features learning from Sentinel-2 data using recurrent-convolutional neural network (R-CNN)," Applied Sciences, vol. 10, no. 1, pp. 238, 2019.
[24] J. A. Kim, J. Y. Sung, and S. H. Park, "Comparison of Faster-RCNN, YOLO, and SSD for real-time vehicle type recognition," in 2020 IEEE International Conference on Consumer Electronics-Asia, 2020, pp. 1-4.
[25] L. I. Yundong, D. O. N. G. Han, L. I. Hongguang, X. Zhang, B. Zhang, and X. I. A. O. Zhifeng, "Multi-block SSD based on small object detection for UAV railway scene surveillance," Chinese Journal of Aeronautics, vol. 33, no. 6, pp. 1747-1755, 2020.
[26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," in Computer Vision - ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I, 2016, pp. 21-37, Springer International Publishing.
[27] Q. Yin, W. Yang, M. Ran, and S. Wang, "FD-SSD: An Improved SSD Object Detection Algorithm Based on Feature Fusion and Dilated Convolution," Signal Processing: Image Communication, vol. 98, p. 116402, 2021.
[28] K. R. Akshatha, A. K. Karunakar, S. B. Shenoy, A. K. Pai, N. H. Nagaraj, and S. S. Rohatgi, "Human Detection in Aerial Thermal Images Using Faster R-CNN and SSD Algorithms," Electronics, vol. 11, no. 7, p. 1151, 2022.
[29] F. Yizhou, C. Junyan, X. Tianmin, C. Rongfeng, L. Xinyu, T. Yongqing, and L. Xiaochun, "Application of the SSD Algorithm in a People Flow Monitoring System," in 2019 15th International Conference on Computational Intelligence and Security (CIS), 2019, pp. 341-344, IEEE.
[30] S. Kim and H. Kim, "Zero-Center Fixed-Point Quantization with Iterative Retraining for Deep Convolutional Neural Network-Based Object Detectors," IEEE Access, vol. 9, pp. 20828-20839, 2021.
[31] F. F. dos Santos, P. Navaux, L. Carro, and P. Rech, "Impact of reduced precision in the reliability of deep neural networks for object detection," in Proceedings of the 2019 IEEE European Test Symposium, May 2019, pp. 1-6.
[32] G. Lan, L. De Vries, and S. Wang, "Evolving efficient deep neural networks for real-time object recognition," in Proceedings of the 2019 IEEE Symposium Series on Computational Intelligence, December 2019, pp. 2571-2578.
[33] R. Ravindran, M. J. Santora, and M. M. Jamali, "Multi-object detection and tracking, based on DNN, for autonomous vehicles: A review," IEEE Sensors Journal, vol. 21, no. 5, pp. 5668-5677, 2020.
[34] Y. Zhu, C. Chen, G. Yan, Y. Guo, and Y. Dong, "AR-Net: Adaptive attention and residual refinement network for copy-move forgery detection," IEEE Transactions on Industrial Informatics, vol. 16, no. 10, pp. 6714-6723, 2020.
[35] B. Wan, Y. Fang, X. Xia, and J. Mei, "Weakly supervised video anomaly detection via center-guided discriminative learning," in Proceedings

of the 2020 IEEE International Conference on Multimedia and Expo, July 2020, pp. 1-6.

[36] S. Yu, C. Wang, Q. Mao, Y. Li, and J. Wu, "Cross-epoch learning for weakly supervised anomaly detection in surveillance videos," IEEE Signal Processing Letters, vol. 28, pp. 2137-2141, 2021.

[37] K. Nalaie, R. Xu, and R. Zheng, "DeepScale: Online Frame Size Adaptation for Multi-object Tracking on Smart Cameras and Edge Servers," in Proceedings of the 2022 IEEE/ACM Seventh International Conference on Internet-of-Things Design and Implementation, May 2022, pp. 67-79.

[38] E. Duman and O. A. Erdem, "Anomaly detection in videos using optical flow and convolutional autoencoder," IEEE Access, vol. 7, pp. 183914-183923, 2019.

[39] J. Salido, V. Lomas, J. Ruiz-Santaquiteria, and O. Deniz, "Automatic handgun detection with deep learning in video surveillance images," Applied Sciences, vol. 11, no. 13, pp. 6085, 2021.

[40] A. Castillo, S. Tabik, F. Pérez, R. Olmos, and F. Herrera, "Brightness guided preprocessing for automatic cold steel weapon detection in surveillance videos with deep learning," Neurocomputing, vol. 330, pp. 151-161, 2019

[41] M. Nalamati, A. Kapoor, M. Saqib, N. Sharma, and M. Blumenstein, "Drone detection in long-range surveillance videos," in 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1-6, 2019.

[42] C. Kwan, D. Gribben, B. Chou, B. Budavari, J. Larkin, A. Rangamani, and R. Etienne-Cummings, "Real-time and deep learning based vehicle detection and classification using pixel-wise code exposure measurements," Electronics, vol. 9, no. 6, pp. 1014, 2020.

[43] R. Solovyev, W. Wang, and T. Gabruseva, "Weighted boxes fusion: Ensembling boxes from different object detection models," Image and Vision Computing, vol. 107, pp. 104117, 2021.

[44] K. Boudjit and N. Ramzan, "Human detection based on deep learning YOLO-v2 for real-time UAV applications," Journal of Experimental & Theoretical Artificial Intelligence, vol. 34, no. 3, pp. 527-544, 2022.

[45] M. T. Bhatti, M. G. Khan, M. Aslam, and M. J. Fiaz, "Weapon detection in real-time cctv videos using deep learning," IEEE Access, vol. 9, pp. 34366-34382, 2021.

[46] M. Rashid, M. A. Khan, M. Alhaisoni, S. H. Wang, S. R. Naqvi, A. Rehman, and T. Saba, "A sustainable deep learning framework for object recognition using multi-layers deep features fusion and selection," Sustainability, vol. 12, no. 12, pp. 5037, 2020.

[47] M. Krišto, M. Ivasic-Kos, and M. Pobar, "Thermal object detection in difficult weather conditions using YOLO," IEEE Access, vol. 8, pp. 125459-125476, 2020.

[48] S. Lu, B. Wang, H. Wang, L. Chen, M. Linjian, and X. Zhang, "A real-time object detection algorithm for video," Computers & Electrical Engineering, vol. 77, pp. 398-408, 2019.

[49] K. Fu, Q. Zhao, I. Y. H. Gu, and J. Yang, "Deepside: A general deep framework for salient object detection," Neurocomputing, vol. 356, pp. 69-82, 2019.

[50] T. Ahmad, Y. Ma, M. Yahya, B. Ahmad, S. Nazir, and A. U. Haq, "Object detection through modified YOLO neural network," Scientific Programming, 2020, pp. 1-10, 2020.

[51] Li, C. J., Qu, Z., Wang, S. Y., & Liu, L. (2021). A method of cross-layer fusion multi-object detection and recognition based on improved faster R-CNN model in complex traffic environment. Pattern Recognition Letters, 145, 127-134.

[52] Shorfuzzaman, M., Hossain, M. S., & Alhamid, M. F. (2021). Towards the sustainable development of smart cities through mass video surveillance: A response to the COVID-19 pandemic. Sustainable Cities and Society, 64, 102582.

[53] Yang, X., Wu, T., Zhang, L., Yang, D., Wang, N., Song, B., & Gao, X. (2019). CNN with spatio-temporal information for fast suspicious object detection and recognition in THz security images. Signal Processing, 160, 202-214.

[54] Chin, T. W., Ding, R., & Marculescu, D. (2019). Adascale: Towards real-time video object detection using adaptive scaling. Proceedings of Machine Learning and Systems, 1, 431-441.

[55] Muhammad, K., Ahmad, J., Lv, Z., (et al., 3 more authors). (2019) Efficient deep CNN-based fire detection and localization in video surveillance applications. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 49 (7), 1419-1434. doi: 10.1109/TSMC.2019.2901260

[56] Chen, Z., Khemmar, R., Decoux, B., Atahouet, A., & Ertaud, J. Y. (2019, July). Real time object detection, tracking, and distance and motion estimation based on deep learning: Application to smart mobility. In 2019 Eighth International

Conference on Emerging Security Technologies (EST), 1-6. doi: 10.1109/EST.2019.8861271

[57] Goyal, A., Anandamurthy, S. B., Dash, P., Acharya, S., Bathla, D., Hicks, D., ... & Ranjan, P. (2020). Automatic border surveillance using machine learning in remote video surveillance systems. In Emerging Trends in Electrical, Communications, and Information Technologies: Proceedings of ICECIT-2018, 751-760. Springer Singapore. doi: 10.1007/978-981-15-0601-7_74

[58] Zhang, D., Maei, H., Wang, X., & Wang, Y. F. (2017). Deep reinforcement learning for visual object tracking in videos. arXiv preprint arXiv:1701.08936.

[59] Zhang, Y., Wang, C., Wang, X., Zeng, W., & Liu, W. (2021). Fairmot: On the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision, 129, 3069-3087. doi: 10.1007/s11263-021-02050-9

[60] Papakis, I., Sarkar, A., & Karpatne, A. (2020). Gcnnmatch: Graph convolutional neural networks for multi-object tracking via sinkhorn normalization. arXiv preprint arXiv:2010.00067.

[61] W. Cao, J. Yuan, Z. He, Z. Zhang, and Z. He, "Fast Deep Neural Networks with Knowledge Guided Training and Predicted Regions of Interests for Real-Time Video Object Detection," IEEE Access, vol. 6, pp. 8990-8999, 2018.

[62] W. Gan, S. Wang, X. Lei, M. S. Lee, and C. C. J. Kuo, "Online CNN-Based Multiple Object Tracking with Enhanced Model Updates and Identity Association," Signal Processing: Image Communication, vol. 66, pp. 95-102, 2018.

[63] P. Zhang, T. Zhuo, W. Huang, K. Chen, and M. Kankanhalli, "Online Object Tracking Based on CNN with Spatial-Temporal Saliency Guided Sampling," Neurocomputing, vol. 257, pp. 115-127, 2017.

[64] Z. Hu, D. Yang, K. Zhang, and Z. Chen, "Object Tracking in Satellite Videos Based on Convolutional Regression Network with Appearance and Motion Features," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 783-793, 2020.

[65] M. Elhoseny, "Multi-Object Detection and Tracking (MODT) Machine Learning Model for Real-Time Video Surveillance Systems," Circuits, Systems, and Signal Processing, vol. 39, pp. 611-630, 2020.

[66] J. Zhang, L. Zi, Y. Hou, M. Wang, W. Jiang, and D. Deng, "A Deep Learning-Based Approach to Enable Action Recognition for Construction Equipment," Advances in Civil Engineering, vol. 2020, pp. 1-14, 2020.

[67] A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Action Recognition Using Optimized Deep Autoencoder and CNN for Surveillance Data Streams of Non-Stationary Environments," Future Generation Computer Systems, vol. 96, pp. 386-397, 2019.

[68] N. Dua, S. N. Singh, and V. B. Semwal, "Multi-Input CNN-GRU Based Human Activity Recognition Using Wearable Sensors," Computing, vol. 103, pp. 1461-1478, 2021.

[69] N. Jaouedi, N. Boujnah, and M. S. Bouhlel, "A New Hybrid Deep Learning Model for Human Action Recognition," Journal of King Saud University-Computer and Information Sciences, vol. 32, no. 4, pp. 447-453, 2020.

[70] M. A. Khan, K. Javed, S. A. Khan, T. Saba, U. Habib, J. A. Khan, and A. A. Abbasi, "Human Action Recognition Using Fusion of Multiview and Deep Features: An Application to Video Surveillance," Multimedia Tools and Applications, pp. 1-27, 2020.

[71] S. Wan, L. Qi, X. Xu, C. Tong, and Z. Gu, "Deep learning models for real-time human activity recognition with smartphones," Mobile Networks and Applications, vol. 25, pp. 743-755, 2020.

[72] H. Mliki, F. Bouhlel, and M. Hammami, "Human activity recognition from UAV-captured video sequences," Pattern Recognition, vol. 100, pp. 107140, 2020.

[73] K. Xia, J. Huang, and H. Wang, "LSTM-CNN architecture for human activity recognition," IEEE Access, vol. 8, pp. 56855-56866, 2020.

[74] Y. Xing, C. Lv, H. Wang, D. Cao, E. Velenis, and F. Y. Wang, "Driver activity recognition for intelligent vehicles: A deep learning approach," IEEE Transactions on Vehicular Technology, vol. 68, no. 6, pp. 5379-5390, 2019.

[75] Y. Meng, C. C. Lin, R. Panda, P. Sattigeri, L. Karlinsky, A. Oliva, et al., "Ar-net: Adaptive frame resolution for efficient action recognition," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII, pp. 86-104, Springer International Publishing, 2020.

[76] S. Jha, C. Seo, E. Yang, and G. P. Joshi, "Real time object detection and tracking system for video surveillance system," Multimedia Tools and Applications, pp. 1-15, 2020, doi: 10.1007/s11042-020-09749.