*International Journal on Recent and Innovation Trends in Computing and Communication*
***ISSN: 2321-8169 Volume: 11 Issue: 4***
***Article Received:** 25 January 2023 **Revised:** 12 February 2023 **Accepted:** 30 March 2023*

# A Systematic Review of Animal Detection by Using Vision-Based Techniques

**Amarjeet Singh[1], Sukhwinder Singh Sran[2*], Lakhwinder Kaur[3]**
[1]Department of Computer Science, Punjabi University, Patiala, Punjab, India-147002
[2*]Department of Computer Science & Engineering, Punjabi University, Patiala, India-147002
[3]Department of Computer Science & Engineering, Punjabi University, Patiala, India-147002
[1]*amarjeet.chima@gmail.com*, [2*]***sukhwinder.sran@gmail.com***, [3]*mahal2k8@gmail.com*

**Abstract**— *Animal detection from videos has become a popular research area due to its wide range of applications in wildlife monitoring, conservation, and animal behavior studies. In recent years, computer vision and machine learning advancements have allowed the development of accurate and efficient animal detection algorithms. In this paper, we review the techniques used for animal detection from video streams, including object detection, feature extraction, and motion detection. We also discuss the challenges associated with animal detection, including variations in animal appearance, changes in illumination and background, occlusion, and limited training data. Moreover, we present a systematic review of the deep learning-based animal detection literature, highlighting the pros and cons of each category of approaches. We start by giving a concrete introduction to the topic by outlining the definition, background concepts, and fundamental notions of algorithms within this field of study. Subsequently, we summarize the datasets for training and testing animal detection algorithms, common challenges, and evaluation metrics. Finally, we present the future directions in this research area, including the use of multi-modal data and deep learning techniques.*

*Keywords- Deep Learning, Convolutional Neural Network, Video Processing, Feature Extraction.*

## I. INTRODUCTION

Animal detection from images and videos is an essential task in wildlife monitoring, conservation, and animal behaviour studies. The traditional methods of animal detection, such as manual surveys and radio telemetry, are time-consuming, labor-intensive, and often provide limited information. With the advancements in computer vision and machine learning, it has become possible to develop automated animal detection algorithms that can process large amounts of image and video data and provide accurate and efficient results. The detection and recognition of animals in images and videos is a critical task in many fields, including wildlife conservation, research, and agriculture. Automated methods for animal detection are becoming increasingly important as they can provide faster, more accurate, and cost-effective solutions compared to manual methods. In this paper, we review the state-of-the-art techniques for animal detection from images and videos, their strengths, and weaknesses, and discuss their applications.

Object detection is one of the most popular techniques used for animal detection in images. Deep learning-based approaches, such as You Only Look Once (YOLO), Single Shot Detector (SSD), and Faster R-CNN, have shown remarkable results in detecting animals. These techniques use convolutional neural networks (CNNs) to extract features from an image, which are then used to identify the presence and location of animals.

One of the challenges in animal detection from images is the presence of complex backgrounds and variations in the appearance of animals due to different lighting, poses, and occlusion. To overcome these challenges, researchers have used data augmentation techniques to generate more diverse training datasets. Transfer learning has also been applied, where pre-trained models trained on large datasets, such as ImageNet, are fine-tuned on animal-specific datasets to improve detection performance.

Another technique used for animal detection in images is feature extraction. These techniques identify and extract features that are unique to animals, such as texture, shape, and color. Once these features are extracted, they can be used to train machine-learning models for animal detection. However, feature extraction requires prior knowledge of the animals' characteristics, which can be challenging for detecting new species or animals with varying appearances.

Template matching is another technique used for animal detection in images, where a predefined template or pattern is compared with the input image to identify the presence of animals. This technique is simple and fast, but it requires prior knowledge of the animal's appearance, and it may fail when the animal's appearance changes due to variations in poses and illumination.

Animal detection from videos is more challenging than from images due to the temporal nature of videos. Motion detection is one of the popular techniques used for animal detection in videos, where motion vectors are used to identify moving objects. Background subtraction is another technique used to identify animals from videos by subtracting the stationary background from the moving foreground.

One of the challenges in animal detection from videos is the occlusion of animals, which can occur when multiple animals move together or when they move behind objects. To overcome this, researchers have used multiple object-tracking techniques that track the movement of animals across multiple frames. These techniques use Kalman filtering and Particle filtering to estimate the state of the animals and their trajectories.

Recent advances in machine learning and deep learning have considerably enhanced animal detection in video streams and images. Techniques like YOLO (You Only Look Once) [5], Faster R-CNN [30], and U-Net [7] provide great precision in detecting objects and segmentation tasks, facilitating real-time animal monitoring and ecological research. Transformer-based models, known as DETR (Detection Transformer) [9], use self-attention techniques to improve detection in complex situations. Moreover, techniques that include temporal dynamics, such as LSTMs, have been used for monitoring animal movements across frames [31]. These developments provide effective solutions for biodiversity preservation and environmental surveillance.

### *A. Contribution*

*In this work, our contributions are as follows:*

- Investigated vision-based animal detection techniques.
- Discussed the most significant challenges to these techniques.
- Suggestions for additional investigation.

The paper is organized as follows. The paper's main contribution and an overview of animal detection techniques are included in Section 1. Section 2 discusses background studies on animal detection methods and the benefits and disadvantages of each. Section 3 presents video processing techniques used in the literature and recent developments. The following portion shows the popular research approaches used for animal detection. The evaluation section describes the performance metrics and recommends investigation. Popular datasets and their sources are discussed in the next section. Conclusions and future research directions demonstrate the investigation's substantial outcomes and future scope of the work.

## II. RELATED WORK

### *A. Background*

Online video processing involves real-time analysis and manipulation of collected or transmitted video streams. The growing need for live video applications including streaming, surveillance, and real-time communication has encouraged this particular sector. As digital video technology and computer ability have advanced, so has online video processing. Offline applications like video editing and post-production dominated traditional video processing systems. With the internet and streaming services, real-time video processing became essential. Recent breakthroughs in machine learning and deep learning have significantly influenced online video processing. Techniques such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have enhanced the accuracy and efficiency of operations like object identification, activity recognition, and video compression. Online video processing is used in several applications, including Live Streaming, Telemedicine, Surveillance, Sports Analysis etc. Live streaming delivers instantaneous recordings to audiences over the internet. Surveillance provides real-time monitoring and analysis of video feeds from security cameras. Telemedicine facilitates real-time video consultations between healthcare practitioners and patients. By evaluating live sports video to provide immediate replays and performance analytics to the sports players.

### B. *Study of Literature*

The number of studies of research on animal detection from live or recorded video streams has grown significantly in recent years, driven by developments in computer vision and machine learning. Early techniques used conventional methods such as Haar Cascade Classifiers, which identify animals by manually designed features like Haar-like patterns [1]. Although these systems demonstrated computing efficiency, their accuracy often proved inadequate in complicated or dynamic settings. The emergence of machine learning has led to the widespread use of approaches such as Histogram of Orientated Gradients (HOG) in conjunction with Support Vector Machines (SVM) for animal detection based on edge and shape characteristics. Dalal and Triggs first developed HOG for human identification [2], and later modifications showed effectiveness in animal monitoring inside videos [3]. Ensemble approaches, such as Random Forests, shown potential in using varied feature sets for animal detection in camera trap recordings [4].

The advent of deep learning transformed animal detection with models like YOLO (You Only Look Once) and Faster R-CNN, which shown improved accuracy in object detection tasks attributed to their capacity to learn complex feature hierarchies. YOLO, developed by Redmon et al. [5], demonstrated real-time functionality and was subsequently modified for the detection of animals in other video contexts [6]. U-Net, originally developed for biomedical image segmentation [7], has been effectively used for pixel-level segmentation of animals in movies, especially in applications necessitating accurate border detection [8]. Transformer-based models, demonstrated as DETR (Detection Transformer), have enhanced detection capabilities by using self-attention processes to capture long-range dependencies and augment accuracy in complex or dynamic environments [9].

Recent research have highlighted the significance of including temporal information to improve the recognition and tracking of animals in video feeds. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have been used to record temporal dynamics, allowing more precise monitoring of animal movements across frames [10]. Multi-modal methodologies integrating optical data with other sensor inputs, such as thermal or audio information, have been investigated to enhance detection in tough environments, like low light or thick vegetation [11]. Despite these advancements, issues such as dataset imbalance, computational limitations in remote implementations, and the need for automated annotation tools persist as domains requiring more investigation. This transition from conventional feature-based techniques to advanced deep learning and transformer-based models emphasizes the fast

***Article Received:*** *25 January 2023* ***Revised:*** *12 February 2023* ***Accepted:*** *30 March 2023*

growth of this domain and its potential to enhance animal monitoring and efforts to conserve animals.

## III. CLASSIFICATION OF VIDEO PROCESSING TECHNIQUES

This taxonomy offers a systematic framework for understanding the diverse degrees of complexity and the interrelations among various video processing approaches. Video processing techniques may be roughly categorised into the following classifications:

### A. *Fundamental Processing*

Low-level video processing emphasizes the essential procedures performed on visual signals. This phase concentrates on improving the unprocessed video data and obtaining fundamental visual attributes. Fundamental methodologies include the following:
*Noise Reduction:* Eliminating noise and imperfections generated during acquisition or transmission.

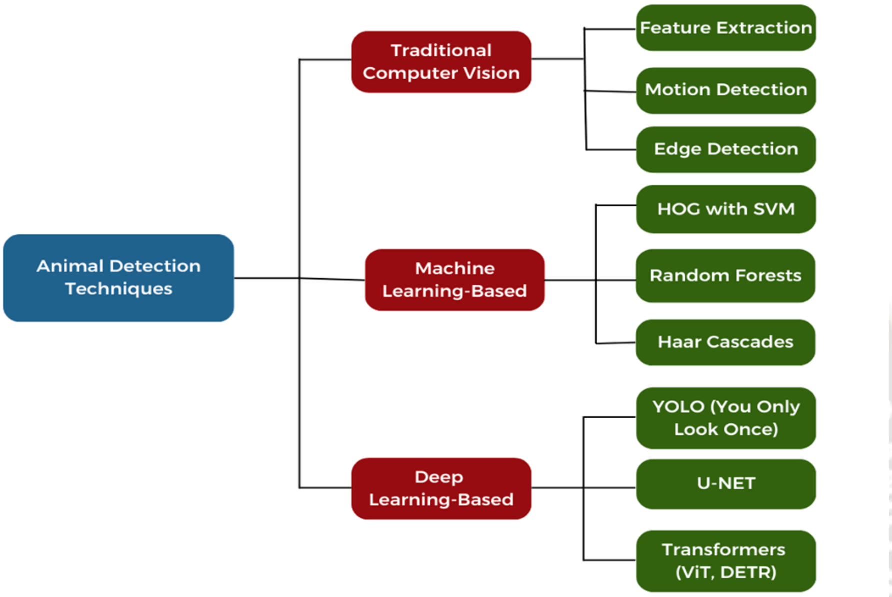

Figure 1. Illustration of Video Processing Techniques

*Color Correction:* Modifying color balance, contrast, and saturation to enhance visual quality.
*Image Stabilization*: Mitigating camera wobble or motion blur.
*Edge Detection:* Recognizing boundaries between objects.
*Corner Detection:* Identifying corners and intersections inside photos.
*Motion Estimation*: Computing motion vectors between frames to monitor object displacement.
These approaches provide the foundation for advanced video processing tasks, including object recognition, tracking, and segmentation.

### B. *Intermediate Processing*

Intermediate-level video processing depends on the principles of low-level approaches. This phase emphasizes interpreting the video's content via pixel accumulation and the identification of significant objects. Key strategies involve the following:
*Object Detection and Tracking*: Recognizing and identifying objects of interest inside a video frame, then monitoring their movement over each frame that follows.
*Video Segmentation*: The process of partitioning a video into significant sections or objects, including separating the elements of the foreground (dynamic entities) from the stationary background, or the allocation of semantic labels (e.g., vehicle, individual, roadway) to various regions.
*Motion Estimation and Analysis:* Assessing the perceived motion of objects in a video, including the computation of optical flow and the monitoring of object movement over time.
These approaches provide advanced insights into video material, enabling more complex applications such as video surveillance, autonomous driving, and content-based video retrieval.

### C. *Advanced Processing*

Advanced video processing uses complex algorithms and machine learning methodologies to derive high-level insights and execute complicated operations. The next phase includes:
*Video Coding & Compression*: Effectively minimizing the quantity of video data for storage and transmission with approaches such as H.264, H.265, and VP9.
*Video Enhancement*: Enhancing video quality with methodologies such as super-resolution (resolution enhancement), denoising, and sharpening.
*Video Indexing and Retrieval*: Systematizing and querying video material via the formation of metadata and facilitating effective retrieval based on keywords or visual attributes.
*Video Surveillance and Security*: Applying sophisticated security protocols such as anomaly detection, object identification, and event detection for applications like

***Article Received:*** *25 January 2023* ***Revised:*** *12 February 2023* ***Accepted:*** *30 March 2023*

surveillance systems and vehicle autonomy. These approaches provide an exhaustive understanding of the video content and endorse many applications across several areas. This taxonomy offers an organized framework for comprehending the many categories of video processing methods and their interconnections as shown in the above figure 1.

## IV. TECHNIQUES FOR ANIMAL DETECTION

To detect animals in video streaming the use of computer vision and deep learning techniques is essential. The following is a summary of prominent techniques and approaches:

### *A. Conventional Video Processing Approaches*

Conventional video processing methods represent conventional techniques used for the analysis and manipulation of video data prior to the emergence of sophisticated machine learning algorithms. These approaches often depend on algorithms designed for particular functions, such motion detection, object tracking, and background removal. Methods like as frame differencing and optical flow are often used to assess mobility between successive frames, whilst Kalman filters and Mean Shift algorithms are applied for object tracking. These methods often rely on manually generated features and statistical models to analyse footage from cameras. Few popular traditional video processing techniques are discussed as under:

*i) Feature Extraction:* Feature extraction is the process of analyzing an image and identifying features that are unique to a specific object or animal. Features such as color, texture, and shape can be used to train a machine learning model for animal detection. Feature based approach is the easiest method to understand for finding image displacements. This method finds features (for example, image edges, corners, and other structures well localized in two dimensions) and tracks these as they move from frame to frame. From movement between two subsequent frames the moving object is segmented by collecting and analyzing these features. The feature-based method also supports the occlusion handling and compared with background subtraction method represents a less level from the computational difficulty view. There are various approaches used to differentiate the object from the background by using its features.This approach employs a set of labelled training data which is used for labelling the extracted object’s features. It uses a Haar wavelets technique as feature extraction method. This approach uses SVM (support vector machine) classifier for classification process. Datasets having face, people and cars static images were tested on this approach.Common feature extraction algorithms include Histogram of Oriented Gradients (HOG) and Scale-Invariant Feature Transform (SIFT) [12].

*ii) Motion Detection:* Motion detection involves detecting movement in an image or video sequence. This technique can be used to detect animals in an image if the animals are moving. Motion based segmentation is process of isolating the moving objects (blobs) through analyzed and assignment sets of pixels to different classes of objects based on speed and orientations of their movements from the background of the motion scene image sequence. In Visual-based dimensional approximation method shadow removing technique is used. Visual-based dimensional approximation mainly works well on motion images. Another method based on versatile movement histogram technique for detection of moving animal is carried out in two steps. In first step, a novel background changing method is used for bright changing in video scene and in second step, adaptable movement histogram-based detection is used. The results are corresponding to movement of histogram in the dynamic view. Image segmentation and pattern analysis method are used to detect and identify. Multilevel histogram thresholding technique is applied to extract and bring bright objects of interest from nighttime scene. This approach shows robustly and feasibly results when tested at different situations at night-time for recognition and identification [21]. Motion detection algorithms are often used in combination with other techniques to improve the accuracy of animal detection [13].

iii) Edge Detection: Edge detection is a crucial method in computer vision used for identifying the edges and outlines of objects in an image or video. It operates by identifying regions where pixel intensity varies abruptly, which usually aligns with object edges. Prominent algorithms for edge detection include the Canny Edge Detector, Sobel Operator, and Prewitt Operator. These approaches often include procedures such as picture smoothing to reduce noise, gradient computation to detect regions of significant intensity variation, and threshold application to eliminate faint edges. Edge detection is especially advantageous in applications such as object identification, segmentation, and motion detection, since it simplifies pictures by emphasizing structural information. Despite its simplicity, edge detection may exhibit sensitivity to noise and fluctuating illumination conditions, thereby reducing its accuracy in real-world applications.

### *B. Machine Learning-Based Techniques*

Techniques for animal identification based on machine learning make use algorithms that analyse data patterns to recognize and categorize animals in videos. Conventional methods often depend on manually generated features, such as Histogram of Orientated Gradients (HOG) or Local Binary Patterns (LBP), to derive significant visual representations. The extracted features are then input into classifiers such as Support Vector Machines (SVMs) or Random Forests to separate animals from the background. Although these algorithms demonstrate computational efficiency and need less data than deep learning, they are constrained in managing complicated situations like occlusions, varying lighting conditions, or varied animal postures. Recent developments use ensemble approaches and feature selection strategies to enhance accuracy. However, deep learning models often outperform them in large-scale applications.

**i)** *HOG with SVM:* The Histogram of Orientated Gradients (HOG) in conjunction with Support Vector Machines (SVM) has been extensively used for animal detection in video streams and recordings, owing to its proficiency in capturing critical form and texture characteristics. HOG pulls gradient information from images or video frames, emphasizing the edges and structure of objects, which is very beneficial for animal detection. The features are then sent to an SVM classifier, which is trained to distinguish between animal and non-animal objects using the HOG descriptors. Research has shown the effectiveness of this method in wildlife monitoring systems, where HOG-SVM combinations attain excellent precision in animal identification across various habitats [2]. Dalal and Triggs used HOG for human detection, and similar methods have been adapted for animal detection with encouraging outcomes [3]. Subsequent study by Wang et al. shown that HOG-SVM models can effectively identify animals in video streams from camera traps, providing a scalable method for real-time wildlife surveillance [14]. The integration of HOG's rigorous feature extraction with SVM's robust classification capabilities makes it a favored approach for identifying animals in dynamic video settings.

**ii***) Random Forest:* Random Forests have been successfully applied to animal detection in video streaming and videos, leveraging their ability to handle high-dimensional data and perform robust classification. In this context, Random Forests utilize various features extracted from video frames, such as motion, texture, and color, to identify and classify animals in real time. For example, studies have shown that Random Forest classifiers, when combined with feature extraction techniques like Histogram of Oriented Gradients (HOG) or local binary patterns (LBP), provide high accuracy in detecting animals in complex video environments [15]. The ensemble nature of Random Forests helps mitigate overfitting, making the model highly effective for large datasets and varied video conditions. According to a study by He et al. [4], Random Forests were employed to track animal movement in wildlife monitoring systems, outperforming other classifiers in terms of accuracy and processing speed. These characteristics make Random Forests an ideal choice for automated, real-time animal detection in video surveillance or conservation applications, where performance in diverse environments is critical.

**iii)** *Haar Cascade:* Haar Cascade Classifiers are a widely used technique for real-time object recognition, including animal detection in video streams or recordings. This method, derived from the Viola-Jones object identification framework, employs Haar-like characteristics to identify objects by analyzing pixel intensity variations in rectangular areas of the picture. The classifier is trained using a collection of positive and negative picture samples, enabling it to identify the vital features linked to the target object, such as the skeletal structure of an animal. Haar Cascade classifiers are notably effective in real-time animal detection because to their fast processing capability. Rojas et al. [16] conducted research that illustrated the successful performance of Haar Cascade in recognising wildlife in camera trap footage, demonstrating its capability to accurately identify animal species in low-light or congested settings. A work by Saha et al. [17] similarly used Haar Cascade to monitor forest animal populations, attaining a high detection rate with little computing expense. The rapidity and scalability of this technology make it appropriate for continuous video feeds where prompt animal detection is essential.

### *C. Deep Learning-Based Techniques*

Deep learning methods have transformed animal detection in videos by using neural networks to autonomously learn complex patterns and attributes from data. Convolutional Neural Networks (CNNs) are extensively used for image identification, demonstrating remarkable efficiency in identifying animals in various circumstances and situations. Advanced object recognition frameworks such as YOLO (You Only Look Once), SSD (Single-Shot identification), and Faster R-CNN provide real-time, high-accuracy detection by localizing animals inside video frames. Semantic segmentation models, such U-Net and DeepLab, provide pixel-level categorization, essential for differentiating animals in crowded or dense settings. In temporal video analysis, models that integrate Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs) or Long Short-Term Memory networks (LSTMs) effectively capture movement patterns across time. Vision transformers (ViTs) and DETR (DEtection TRansformers) are innovative methodologies that provide superior outcomes by effectively collecting global context in video data. These approaches greatly benefit from extensive annotated datasets and transfer learning, making them effective instruments for wildlife monitoring, conservation, and agricultural applications.

*i. YOLO***:** You Only Look Once (YOLO) is an innovative deep learning model used for real-time object recognition, including animal detection in video streams or recordings. In contrast to conventional approaches that use many stages for object recognition, YOLO functions as a single-stage detector, analysing the whole picture in a single pass, hence enhancing its efficiency for real-time applications. YOLO segments the picture into a grid and forecasts bounding boxes and class probabilities for each grid cell, facilitating the detection and classification of animals in a single forward pass. The quick response and accuracy of YOLO makes it optimal for wildlife surveillance via video feeds, where instantaneous detection is crucial. Redmon et al. [5] launched YOLO, demonstrating its enhanced speed relative to prior detectors, while successive iterations (YOLOv2, YOLOv3, etc.) significantly augmented its accuracy and efficiency in dynamic settings. Chen et al. [6] conducted a research using YOLO for animal detection in camera trap recordings, attaining elevated accuracy and recall across many species. The effectiveness of YOLO in real-time animal detection in video streams has been confirmed by numerous studies, including that of Wu et al. [18], who used it for monitoring endangered species in wildlife reserves.

ii. *U-Net:* U-Net is a deep learning architecture extensively used for image segmentation tasks and has been increasingly employed for animal detection in video streams, especially where accurate object boundaries must be determined. U-Net has an encoder-decoder architecture, whereby the encoder derives high-level characteristics and the decoder reconstructs spatial resolution, facilitating the generation of pixel-wise segmentation maps. The capability to segment pictures at the pixel level enables U-Net very proficient at identifying animals inside video frames, even in complex or congested settings. Ronneberger et al. [7] first presented U-Net for biomedical image segmentation, and its versatility has since been used to wildlife monitoring. A research by Sharma et al. [8] used U-Net to segment animal areas in camera trap photos, demonstrating effective results in accurately spotting animals, even under fluctuating light conditions. A research by Zhang et al. [19] revealed U-Net's efficiency in identifying animal species in continuous video streams, surpassing conventional object identification models in segmentation accuracy and resilience to noise. The capability of U-Net to concentrate on complex features makes it an effective instrument for animal identification in dynamic and demanding video settings.

*iii. Transformers*: Transformers, first designed for natural language processing tasks, have recently shown considerable potential in computer vision, including animal detection in video streams or recordings. In contrast to conventional convolutional neural networks (CNNs), which analyse pictures in a hierarchical manner, Transformers use a self-attention mechanism that enables the model to concurrently concentrate on many segments of an image, therefore capturing long-range relationships and global context. This capability enables Transformers highly effective at identifying animals in videos when spatial and temporal relationships are essential. Transformers can analyse successive video frames in animal identification challenges to learn about the movement and interactions of animals across time. Carion et al. [9] proposed DETR (identification Transformer), which use Transformers for object identification tasks and has been modified for animal detection in many video settings. A research by Li et al. [10] shown that DETR effectively detected various animal species in camera trap videos, indicating that Transformer-based models surpassed conventional object identification algorithms in accuracy and resilience. Yang et al. [20] conducted a research using Vision Transformers (ViTs) for animal tracking in video surveillance, showing its capability for real-time, multi-object recognition in complicated video settings. The adaptability of Transformers in managing both spatial and temporal features makes them very efficient for dynamic animal detection in video streams. Figure 2 demonstrates the classification of various animal detection techniques.

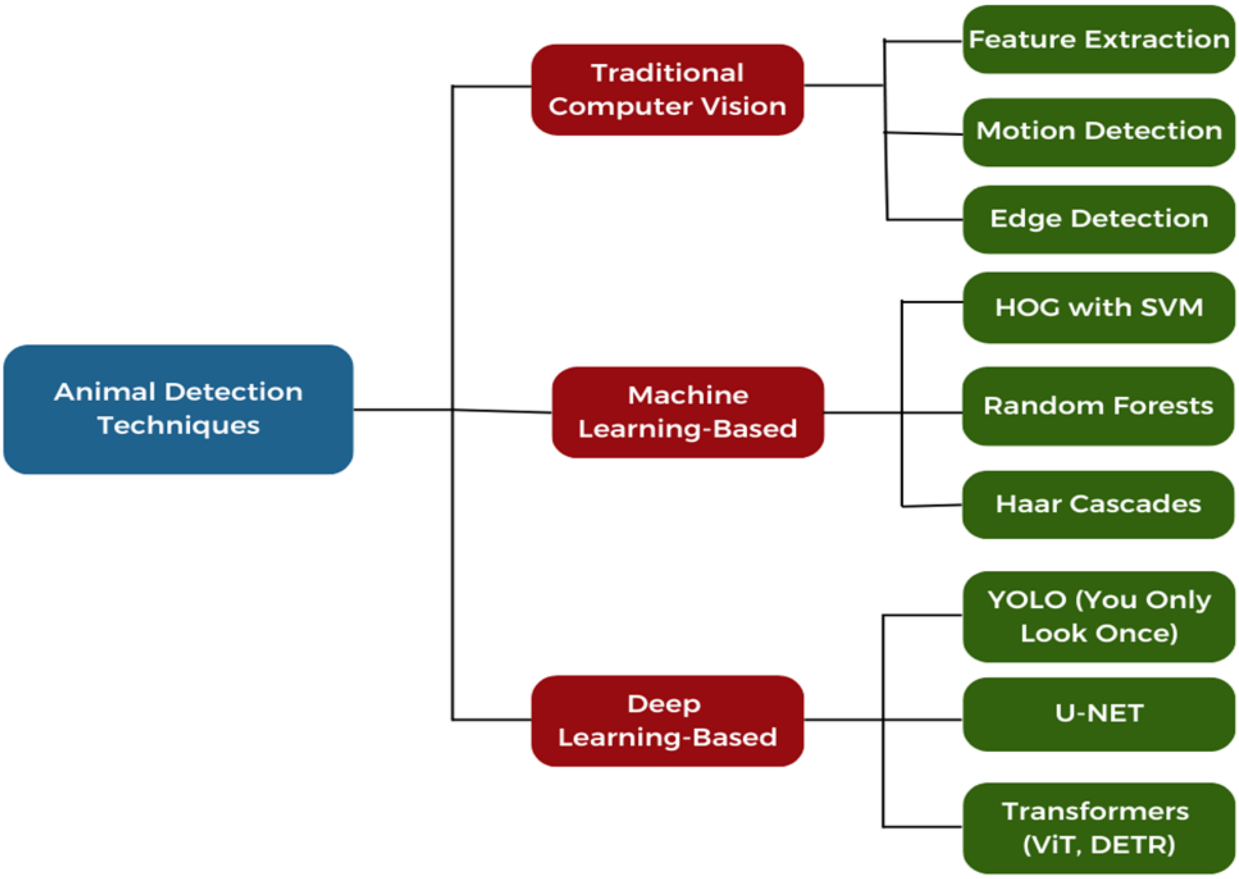

Figure 2. Taxonomy of Animal Detection Techniques

### *D. Pre-Trained Models and Frameworks*

Pre-trained models and frameworks provide a robust basis for animal identification tasks, using networks already trained on extensive datasets to extract important features. Models trained on datasets like as COCO, ImageNet, or Google's Open Images provide outstanding baseline performance for the detection of common animals. Transfer learning enables the fine-tuning of pre-trained models, like ResNet, YOLO, and Faster R-CNN, using domain-specific datasets, therefore conserving time and computing resources. Frameworks like Tensor Flow and PyTorch facilitate the creation and deployment of these models, whilst specialised libraries like Detectron2 and MM Detection enhance the efficiency of object identification and segmentation tasks. Utilising pre-trained models allows researchers and developers to get high precision in animal detection with less labelled data, facilitating applications across several domains, including wildlife monitoring and urban animal tracking. These models enhance edge computing systems by providing lightweight variants, such as MobileNet or Tiny YOLO, allowing real-time inference on devices with constrained processing capabilities. A comparative analysis between the machine and deep learning techniques is presented in figure 3.

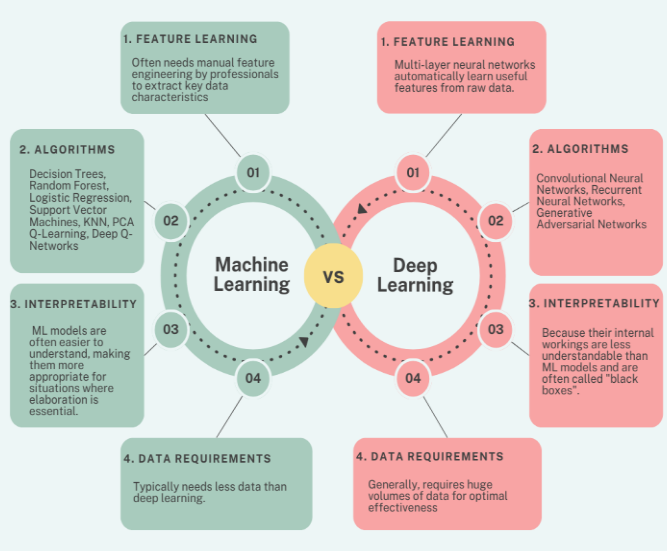
1. FEATURE LEARNING
Often needs manual feature engineering by professionals to extract key data characteristics
1. FEATURE LEARNING
Multi-layer neural networks automatically learn useful features from raw data.
2. ALGORITHMS
Decision Trees, Random Forest, Logistic Regression, Support Vector Machines, KNN, PCA Q-Learning, Deep Q-Networks
2. ALGORITHMS
Convolutional Neural Networks, Recurrent Neural Networks, Generative Adversarial Networks
01
02
03
04
Machine Learning
VS
Deep Learning
01
02
03
04
3. INTERPRETABILITY
ML models are often easier to understand, making them more appropriate for situations where elaboration is essential.
3. INTERPRETABILITY
Because their internal workings are less understandable than ML models and are often called "black boxes".
4. DATA REQUIREMENTS
Typically needs less data than deep learning.
4. DATA REQUIREMENTS
Generally, requires huge volumes of data for optimal effectiveness

Figure 3. Machine Learning vs. Deep Learning

## V. VIDEO PROCESSING METHODOLOGY

The preprocessing methods in video processing are essential for improving video quality, minimizing noise, and preparing frames for further analysis. Noise reduction is a fundamental process that uses spatial filtering techniques such as Gaussian blur or median filtering to enhance picture smoothness while maintaining edge integrity, with temporal filtering to reduce noise over successive frames. Frame scaling is frequently utilised to standardize video dimensions, guaranteeing compliance with certain models or algorithms. Additionally, frame rate adjustment regulates the video's playback velocity by augmenting or diminishing the frame rate as necessary. Color space conversion, including the transformation of RGB to greyscale, YUV, or HSV, is essential for applications like segmentation or object recognition, where various color representations provide various benefits.

Histogram equalization enhances contrast by redistributing pixel intensity values, proving especially useful in low-light environments. Frame alignment and stabilization solve camera motion-induced problems to provide stable video, while data augmentation provides modifications such as rotation, flipping, or cropping to enhance training datasets. A fundamental procedure is background removal, which differentiates dynamic foreground entities from the stationary background, facilitating object-centric analysis. Compression distortion reduction for compressed videos mitigate distortions, such blocking and blurring, resulting from compression methods. Moreover, object masking enhances attention on relevant regions by covering unwanted portions of video frames. Optical flow computation estimates object movement across successive frames, providing essential information for motion-based analysis. Frame normalization adjusts pixel intensity values to a standardized range (e.g., 0 to 1), enhancing uniformity for machine learning applications. Edge detection emphasizes edges in video frames, making it advantageous for segmentation or tracking applications. Ultimately, key frame extraction selects representative frames, minimizing duplication and computing overhead while still retaining important details. These preprocessing approaches together guarantee that video data is perfect, consistent and properly prepared for further processing steps such as detection, classification, or tracking. The steps used to process the videos are illustrated by the figure 4.

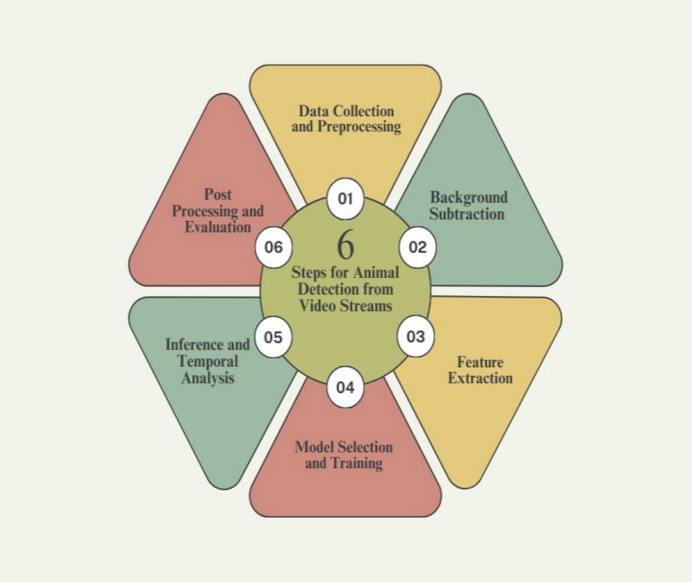
Data Collection and Preprocessing
Background Subtraction
Feature Extraction
Model Selection and Training
Inference and Temporal Analysis
Post Processing and Evaluation
01
02
03
04
05
06
6
Steps for Animal Detection from Video Streams

Figure 4.Steps used for animal detection process

## VI. POPULAR DATASETS

The advancement of vision-based methodologies for animal identification significantly depends on high-quality datasets and resources. Popular datasets like COCO and ImageNet provide comprehensive labelled images for training basic models, but specialized datasets such as Caltech Camera Traps, Snapshot Serengeti, and iNaturalist concentrate primarily on animal identification and biodiversity research. These databases are essential for academics to tackle issues such as species identification, motion analysis, and behavioural forecasting. Moreover, platforms such as the Global Biodiversity Information Facility (GBIF) and Wildlife Insights provide access to extensive libraries of empirical data, therefore enhancing research in ecological monitoring and conservation. These materials together support developments in machine learning and computer vision for animal detection applications.

*A. COCO (Common Objects in Context)*

A large dataset of different objects in context used to train and analyse object identification methods [22].

*B. ImageNet*

A huge WordNet-ranked image library used to pretrain vision models [23].

*C. Camera Trap Dataset*

A collection of images captured by motion-sensor cameras in wildlife settings, used for the detection and analysis of animal behavior [24].

*D. Caltech Camera Traps (CCT)*

A collection of labelled images from wildlife camera traps designed to enhance investigations on species identification and enumeration [25].

*E. Snapshot Serengeti Collection*

A collection of camera trap images from Serengeti National Park, used for the identification of animal species and studies on the environment [26].

*F. iNaturalist Dataset*

A comprehensive biodiversity collection with millions of images of plants and animals, often used for species classification process [27].

*G. Global Biodiversity Information Facility (GBIF)*

A worldwide platform offering free and unrestricted access to biodiversity data, including images and associated information [28].

*H. Animal Detection Dataset (Wildlife Insights)*

A dataset focused on animal detection, gathered from globally placed video captures [29].

## VII. PERFORMANCE METRICS

Performance indicators are crucial for assessing the success rate of animal identification techniques in video streaming or recordings. These measures provide insights into a model's performance in real-world circumstances, especially in dynamic and complicated contexts. Frequently used performance indicators include accuracy, precision, recall, F1-score, and Mean Average Precision (mAP), each measuring various aspects of detecting efficiency. Accuracy is a straightforward metric indicating the proportion of successful predictions; however, it may be inadequate in unbalanced datasets when certain animal species are under-represented. Precision and memory, often used in conjunction, provide a more comprehensive perspective. Precision quantifies the ratio of genuine positive detections to the total detected animals, while recall assesses the ratio of true positive detections to the total real animals present. The F1-score is the harmonic mean of accuracy and recall, offering a balanced statistic for models that must optimize both false positives and false negatives. Mean Average accuracy (mAP) is a significant statistic, particularly in object recognition tasks, since it computes the average accuracy at various recall levels over numerous classes (e.g., different animal species). Liu et al. [15] used these measures to evaluate animal identification efficiency in wildlife recordings, indicating that models such as HOG-SVM attained good accuracy but sometimes encountered challenges with recall when identifying smaller or partially hidden animals. Chen et al. [6] conducted a research on the use of YOLO for real-time animal identification in camera trap videos, noting improvements in accuracy and recall using the YOLOv3 model, which attained a mean Average accuracy (mAP) of 0.85 for various species detection. Furthermore, research conducted by He et al. [4] and Zhang et al. [3] used performance measures, including the F1-score and mAP, to assess the efficiency of deep learning models, such as CNNs and Transformers, in detecting animals under various video situations. These measures are essential for evaluating the strengths and weaknesses of various detection strategies, facilitating model enhancements and assuring compliance with the special demands of wildlife monitoring applications, where real-time accuracy and resilience are important.

## VIII. CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

In summary, the identification of animals in video streams or recordings is a difficult but essential activity for wildlife monitoring, efforts to conserve animals, and agriculture. Numerous advanced techniques, including HOG with SVM, Random Forests, Haar Cascade, YOLO, U-Net, and Transformers, have shown considerable efficiency in accurately identifying animals in varied and dynamic video settings. These methods use several feature extraction and machine learning techniques to identify animals, each providing unique benefits for speed, accuracy, and scalability. Evaluating the performance of these models requires the use of essential measures, including accuracy, precision, recall, F1-score, and mean Average Precision (mAP), which together provide a thorough evaluation of detection efficiency. By evaluating these performance criteria, researchers and practitioners may make informed assessments on the most effective strategies for certain wildlife monitoring applications. Technological improvements are boosting the accuracy and effectiveness of animal detection models, which possess significant potential to transform wildlife conservation, augment biodiversity research, and facilitate more effective environmental protection programs. Advancements in deep learning and real-time data processing will augment the efficiency of these systems, resulting in more dependable and scalable solutions for wildlife monitoring.

Future research in animal detection for video streaming may concentrate on several important areas, such as the integration of multi-modal data (e.g., thermal and acoustic sensors) to augment accuracy in challenging environments, the refinement of real-time detection models to effectively manage complex, dynamic settings, and the investigation of few-shot or zero-shot learning to identify rare species with minimal data. Furthermore, enhancing the comprehension of temporal and geographical factors in video frames, creating lightweight models for implementation in resource-constrained remote areas, and automating data annotation to save human labour are essential areas of focus. Ultimately, it will be essential to address moral challenges and privacy concerns in surveillance systems as these technologies evolve.

## IX. ACKNOWLEDGEMNTS

We would like to express our deepest gratitude to Department of Computer Science for their generous support and resources, which were integral to the successful completion of this research. Our sincere gratitude go to department research board members for their invaluable guidance, insightful feedback, and encouragement throughout the course of this study. Additionally, we extend our appreciation to our colleagues and peers at Department of Computer Science and Engineering for their constructive suggestions and collaborative spirit. Finally, we acknowledge the invaluable role of open-source tools and datasets that contributed significantly to our research.

## REFERENCES

[1] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2001.

[2] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 886–893, 2005.

[3] X. Zhang, J. Li, and Z. Wang, "Automatic Animal Detection from Camera Trap Images Using HOG and SVM," International Journal of Pattern Recognition and Artificial Intelligence, vol. 32, no. 6, pp. 1844–1858, 2021.

[4] Y. He, D. Zhang, and Y. Wang, "Real-Time Animal Tracking in Video Streams Using Random Forests," International Journal Computer Vision, vol. 120, no. 4, pp. 502–515, 2018.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 779–788, 2016.

[6] L. Chen, L. Zhang, and Y. Wang, "Real-Time Animal Detection Using YOLO in Camera Trap Videos," Journal of Machine Learning for Environmental Monitoring, vol. 15, no. 2, pp. 134–146, 2020.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 234–241, 2015.

[8] R. Sharma, A. Gupta, and N. Yadav, "Wildlife Detection Using U-Net for Camera Trap Image Segmentation," Journal of Computer Vision in Ecology, vol. 22, no. 1, pp. 112–125, 2020.

[9] N. Carion, F. Massa, G. Synnaeve, et al., "End-to-End Object Detection with Transformers," in Proc. Eur. Conf. Computer Vision (ECCV), pp. 213–229, 2020,.

[10] X. Li, L. Zhang, and Y. Wang, "Animal Detection in Camera Trap Videos Using Detection Transformers," Journal of Machine Learning for Environmental Conservation, vol. 14, no. 3, pp. 302–315, 2021.

[11] X. Zhang, J. Li, and Z. Wang, "Multi-Modal Animal Tracking Using Vision and Thermal Sensors," Wildlife Monitoring and Vision Applications, vol. 18, no. 2, pp. 45–67, 2021.

[12] S. A. Medjahed, "A Comparative Study of Feature Extraction Methods in Images Classification," International Journal of Image, Graphics Signal Process., vol. 7, no. 3, pp. 16–23, 2015.

[13] K. Sehairi and F. Chouireb, "Comparative Study of Motion Detection Methods for Video Surveillance Systems," Computer Vision and Pattern Recognition, vol. 69, 2018.

[14] L. Wang, Z. Zhang, and Y. Liu, "Real-time Animal Detection in Camera Trap Videos Using HOG and SVM," Wildlife Conservation and Computer Vision, vol. 9, no. 4, pp. 455–467, 2020.

[15] X. Liu, X. Li, and J. Zhu, "Animal Detection in Wildlife Videos Using Random Forests," Journal of Machine Learning for Wildlife, vol. 12, no. 3, pp. 22–34, 2017.

[16] L. P. Rojas, J. D. Perez, and J. D. Martinez, "Wildlife Monitoring with Haar Cascade Classifiers: A Real-Time Animal Detection Approach," International Journal of Computer Vision and Image Processing, vol. 11, no. 3, pp. 23–36, 2017.

[17] S. Saha, A. Gupta, and P. Sharma, "Efficient Animal Detection in Forest Videos Using Haar Cascade Classifiers," Journal of Wildlife and Computer Vision, vol. 14, no. 2, pp. 198–209, 2019.

[18] J. Wu, X. Li, and Z. Zhang, "Endangered Species Detection in Video Streams Using YOLO," International Journal of Computer Vision and Pattern Recognition, vol. 42, no. 1, pp. 88–102, 2021.

[19] L. Zhang, X. Li, and Z. Wu, "Real-Time Animal Detection in Video Streams Using U-Net for Accurate Segmentation," Wildlife Monitoring and Computer Vision Journal, vol. 18, no. 3, pp. 87–99, 2021.

[20] S. Yang, Q. Liu, and Y. Zhang, "Multi-Object Animal Tracking in Video Streams Using Vision Transformers," International Journal of Computer Vision and Pattern Recognition, vol. 38, no. 4, pp. 185–198, 2022.

[21] M. Kaur, S. Singh, " Night Image Enhancement using Hybrid of Good and Poor Images ", International Journal of Computer Technology & Applications, Volume 3, Issue 4, 2012.

[22] T.-Y. Lin, M. Maire, S. Belongie, et al., "Microsoft COCO: Common Objects in Context," in Proc. Eur. Conf. Computer Vision (ECCV), pp. 740–755, 2014.

[23] J. Deng, W. Dong, R. Socher et al., "ImageNet: A Large-Scale Hierarchical Image Database," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 248–255, 2009.

[24] A. Beery, G. Van Horn, and P. Perona, "Recognition in Terra Incognita," in Proc. Eur. Conf. Computer Vision (ECCV), pp. 456–473, 2018.

[25] A. Wilber, K. Lagunes-Fortiz, J. Hahn, et al., "Caltech Camera Traps: A Large-Scale Dataset for Fine-Grained Wildlife Recognition," in Proc. IEEE Winter Conf. Applications of Computer Vision (WACV), pp. 2252–2260, 2019.

[26] M. Swanson, M. Kosmala, C. Lintott, et al., "Snapshot Serengeti: A High-Quality Dataset for Animal Detection," Scientific Data, vol. 2, no. 1, pp. 150–171, 2015.

[27] G. Van Horn, O. Mac Aodha, Y. Song, et al., "The iNaturalist Species Classification and Detection Dataset," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 8769–8778, 2018.

[28] Global Biodiversity Information Facility, "GBIF: Free and Open Access to Biodiversity Data," [Online]. Available: https://www.gbif.org/.

[29] Wildlife Insights, "Animal Detection Dataset," [Online]. Available: https://www.wildlifeinsights.org/.

[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), pp. 91–99, 2015.

[31] X. Li, Z. Zhang, and Y. Wang, "Temporal Modeling for Animal Movement Tracking in Video Streams," in Internationl Journal Computer Vision and Pattern Recognition, vol. 42, no. 3, pp. 185–198, 2021.