

**Multi-Object Tracking and Vision Transformer
Enhancements for Real-Time Cow Monitoring: A
Review and Implementation Study**

A DISSERTATION

SUBMITTED IN PARTIAL FULFILMENT OF REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY
IN
DATA SCIENCE

Submitted by:
SHASHANK GEHLOT
2K23/DSC/16

Under the supervision of

Dr. Rahul
(Assistant Professor)



DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi – 110042

MAY 2025

DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi - 110042

DECLARATION

I hereby affirm that I have completed the thesis entitled "**Multi-Object Tracking and Vision Transformer Enhancements for Real-Time Cow Monitoring: A Review and Implementation Study**" during the year 2024, under the guidance of Dr. Rahul from the Department of Software Engineering at Delhi Technological University, Delhi, as part of the requirements for the partial fulfilment of the M.Tech. degree program offered by the institution. Furthermore, I attest that this thesis is the result of my individual effort and has not been submitted to any other university for any degree award.

Date: 01/07/2025

Place: Delhi

Shashank Gehlot

(2K23/DSC/16)

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor

Signature of External Examiner

DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi - 110042

CERTIFICATE

This is to confirm that Shashank Gehlot (2K23/DSC/16) completed the thesis "**Multi-Object Tracking and Vision Transformer Enhancements for Real-Time Cow Monitoring: A Review and Implementation Study**" under my guidance in partial fulfilment of the MASTER OF TECHNOLOGY degree in Data Science at DELHI TECHNOLOGICAL UNIVERSITY, NEW DELHI. To the best of my knowledge this work has not been submitted in part or full for any other Degree to this University or elsewhere.

Dr. Rahul

(Asst. Professor)

**Department of Software
Engineering, DTU**

Abstract

In recent years, precision livestock farming has emerged as a vital area of innovation, aiming to enhance the health, welfare, and productivity of animals through advanced technologies. This thesis investigates deep learning-based techniques for cow detection, tracking, and behavioral monitoring by integrating insights from a comprehensive literature review and a practical implementation. The review explores the evolution of object detection and tracking models, such as YOLO, R-CNN, and Vision Transformers, evaluating their applicability under varying farm conditions including occlusion, lighting challenges, and dense cattle populations. Building upon this foundation, the thesis proposes a novel hybrid system—YOLOv8s augmented with Coordinate Attention (CA) and integrated with DeepSORT and a Vision Transformer (ViT)—designed for accurate cow face detection and re-identification. The model was trained and tested on a custom dataset collected from a farm in Uttar Pradesh, India, comprising over 400 hours of video. Evaluation metrics show that the proposed system achieved an **IDF1 score of 92.5%**, **MOTA of 88.4%**, and **MOTP of 97.2%**, outperforming traditional methods such as ByteTrack, BoT-SORT, and DeepSORT. Additionally, the system demonstrated a 50% reduction in identity switching and a 20% improvement in processing time. These findings underscore the potential of the proposed approach to serve as a reliable solution for real-time livestock monitoring, contributing to smarter and more efficient dairy farm management.

ACKNOWLEDGEMENT

I am grateful to Dr. Rahul (Asst. Professor, Dept. of Software Engineering) and all of the Department of Software Engineering faculty members at DTU. They all gave us a lot of help and advice for the project.

I'd also want to thank the University for providing us with the laboratories, infrastructure, testing facilities, and environment that allowed us to continue working without interruption.

I'd also like to thank our lab assistants, seniors, and peer group for their aid and knowledge on a variety of subjects.

Shashank Gehlot

2K23/DSC/16

TABLE OF CONTENTS

Candidate Declaration	2
Certificate by the Supervisor(s)	3
Abstract	4
Acknowledgement	5
Table of Contents	6
List of Tables	8
List of Figures	9
List of Abbreviations	10
CHAPTER 1: INTRODUCTION	11-23
1.1 Overview	11
1.2 Background	12
1.2.1 Evolution of Livestock Monitoring Techniques	13
1.2.2 Importance of Deep Learning in Precision Farming	14
1.2.3 Limitations of Conventional Approaches	15
1.3 Problem Statement	16
1.4 Motivation	17
1.5 Contributions of the Thesis	18
1.5.1 Contribution 1: Review of State-of-the-Art Techniques	18
1.5.2 Contribution 2: YOLOv8s-CA + DeepSORT-ViT Implementation	19
1.6 Thesis Organization	20
1.7 Summary	22
CHAPTER 2: LITERATURE SURVEY	24-30
2.1 Overview	24
2.2 Review of Existing Detection & Tracking Techniques	25
2.3 Summary of Reviewed Work	27
2.4 Research Gaps	29
CHAPTER 3: RESEARCH OBJECTIVES	31-33
3.1 Overview	31
3.2 Research Questions	32
3.3 Summary	33
CHAPTER 4: METHODOLOGY	34-40
4.1 Overview	34

4.2 Proposed Model Architecture	35
4.2.1 YOLOv8s-CA: Enhanced Cow Detection	36
4.2.2 DeepSORT-ViT: Improved Re-identification	37
4.3 Data Collection and Dataset Preparation	37
4.4 Model Training and Implementation	38
4.5 Model Evaluation Metrics	39
4.6 Summary	40
CHAPTER 5: RESULTS AND DISCUSSION	41–46
5.1 Overview	41
5.2 Performance Analysis of the Proposed Model	42
5.3 Comparison with Existing Algorithms	43
5.4 Discussion	44
5.5 Summary	46
CHAPTER 6: CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT	47–48
6.1 Conclusion	47
6.2 Future Scope	48
6.3 Social Impact	48
REFERENCES	49

LIST OF TABLES

Table No.	Table Title	Pg No.
Table I	Summary of studies selected for review	27
Table II	Comparison of different tracking methods	44

LIST OF FIGURES

Figure No.	Figure Title	Page No.
Fig 1	Architecture of YOLO (Basic)	13
Fig 2	Review of state of the art techniques	21
Fig 3	Dataset distribution in cow detection studies	28
Fig 4	YOLOv8s-CA+DeepSort-ViT Model For Tracking	34
Fig 5	.Data distribution	37
Fig 6	Comparison of models training time	38
Fig 7	Accuracy metrics of model	43
Fig 8	Comparison Chart Of Tracking Methods	45

LIST OF ABBREVIATIONS

Abbreviation	Long Form
CA	Coordinate Attention
CNN	Convolutional Neural Network
FPS	Frames Per Second
IDF1	Identification F1 Score
MOTA	Multi-Object Tracking Accuracy
YOLO	You Only Look Once
ViT	Vision Transformer
SORT	Simple Online and Realtime Tracking
R-CNN	Region-Based Convolutional Neural Network
MOTP	Multi-Object Tracking Precision

CHAPTER 1: INTRODUCTION

1.1 OVERVIEW

The agriculture and livestock sectors are undergoing a significant transformation with the integration of modern technologies, particularly artificial intelligence (AI), computer vision, and deep learning. One of the most important and time-consuming parts of managing livestock is keeping a close eye on the behavior and health of the animals, particularly in large-scale dairy systems. In addition to using a lot of resources, manual supervision increases the possibility of human mistake, postponed actions, and missed abnormalities [4]. Automating livestock surveillance through sophisticated object detection and tracking systems has been the subject of an expanding corpus of research in recent years. Among these, cow behavior analysis and detection have become popular research topics. The best possible animal care, prompt detection of health problems, and increased farm productivity all depend on advances. In order to overcome current constraints in real-world settings, this thesis proposes a unique implementation based on YOLOv8s-CA and DeepSORT-ViT along with an extensive evaluation of contemporary detection and tracking techniques.

1.2 BACKGROUND

As the global industry of livestock tends to be data-oriented, the transition to AI-oriented technologies has changed how animals are observed in the farms. Specifically, detection and tracking of cows are a crucial task in order to enhance dairy production, detect abnormalities early on, and keep the welfare of animals under control. In open or semi-structured environments, real-time monitoring however creates several problems like background clutter, occlusion, variation in light intensity and similarity between animals [7].

For these, there have been computer vision systems based on the deep learning technology as a promising alternative to the RFID-based or manual tracking systems. The biggest advancement in this area has been the usage of YOLO (You only look once) object detection architecture and its variants, which are real-time with high accuracies.

1.2.1 Evolution of Detection and Tracking in Livestock Farming

The conventional animal monitoring techniques involved manual observation or wearable sensors (such as GPS collars or RFID tag). Those techniques were effective to some extent but

at the same time they had definite disadvantages. sensor maintenance, suffering of animals, and the inability to scale.

The invention of deep convolutional neural networks (CNNs) brought vision-based detection of livestock from camera streams. Initial works used such architectures as R-CNN, SSD, and YOLOv3 for cow detection. Detection got relatively accurate though tracking within the real time was a problem particularly under occlusion and appearance variation. Such limitations was the driving factor behind the integration of MOT pipelines such as DeepSORT [5].

In the last few years, newer variants such as YOLOv5, YOLOv7, and YOLOv8 introduced architectural improvements, allowing deployment for the edge device with less parameters. The inclusion of attention mechanisms such as Coordinate Attention (CA) made object localization in complex scenes even better[8]. At the same time, DeepSORT did not only leverage appearance-based and motion-based cues to maintain tracking across frames consistent but also included the idea of ViT, delivering contextual understanding.

1.2.1.1 YOLO Architecture and Mathematical Formulation

The **YOLO (You Only Look Once)** model is a **single-stage object detector**, meaning it predicts bounding boxes and class labels in a single forward pass. Unlike two-stage detectors that rely on region proposals followed by classification, YOLO[16] is designed for **real-time processing** and is especially effective in resource-constrained environments like farms, where latency and simplicity are critical.

The core idea of YOLO is to divide the input image into a grid of fixed dimensions and let each cell predict a certain number of bounding boxes, along with the **objectness score** and **class probabilities**[19]. The latest versions, such as **YOLOv8**, build on this foundation using improved architecture, better loss functions, and more efficient training techniques

$$\text{Confidence} = \text{Pr}(\text{Object}) \cdot \text{IoU}_{\text{pred}}^{\text{truth}}$$

Where IoU is the **Intersection over Union** between predicted and ground truth box:

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

◆ YOLO Loss Function

The overall YOLO loss combines three components:

$$\mathcal{L}_{\text{YOLO}} = \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbf{1}_{ij}^{\text{obj}} [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2] + \dots$$

This includes localization loss, confidence loss, and classification loss. In YOLOv8, the loss is optimized using **CIoU (Complete IoU)** for better bounding box regression:

$$\text{CIoU} = \text{IoU} - \frac{\rho^2(\mathbf{b}, \mathbf{b}^{gt})}{c^2} - \alpha v$$

Where:

- ρ^2 is the squared Euclidean distance between center points
- c is the diagonal length of the smallest enclosing box
- v measures aspect ratio similarity

◆ Coordinate Attention (CA) in YOLOv8s-CA

YOLOv8s-CA integrates **Coordinate Attention**, which enhances feature maps by encoding precise location information and channel relationships. CA refines convolutional features

$\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ into two attention maps[6]:

$$f_x = \text{AvgPool}_y(X), \quad f_y = \text{AvgPool}_x(X)$$

Then, a shared 1D convolution and sigmoid activation are used to produce spatially-aware weights which modulate XXX, improving detection under occlusion.

- DeepSORT and Vision Transformers for Tracking

Once cows are detected in individual frames, **tracking algorithms** assign a unique ID and follow each cow across frames. DeepSORT improves upon the original SORT by adding appearance descriptors extracted from a CNN-based Re-ID model[5].

The Kalman filter models motion, and **Hungarian algorithm** solves the data association problem:

$$\text{Cost}_{ij} = \lambda_1 \cdot \text{IoU}_{ij} + \lambda_2 \cdot \text{cosine_distance}(f_i, f_j)$$

Where:

- f_i, f_j are the feature embeddings of detections
- λ_1, λ_2 are weights for motion and appearance cues

To overcome appearance similarity among cows, **ViT (Vision Transformer)** embeddings replace CNN-based descriptors. The self-attention mechanism in ViT enables better context-aware identity preservation, reducing ID switches.

1.2.2 Importance of Deep Learning in Precision Farming

The use of deep learning has been revolutionary, especially through Convolutional Neural Networks (CNNs) and object detection models, in the computer vision tasks in agriculture. Being able to automatically extract spatial patterns and complicated visual elements, this technology is perfect for cow detection and tracking[16]. Object detection applications for real time have performed well with algorithms such as YOLO (You Only Look Once), Faster R-CNN and SSD.

Explicitly, in livestock monitoring, these models are trained to detect cows in video feeds with a high level of accuracy, despite having several cows occupy the recording at the same time.

This allows constant observation without the weariness and prejudice that typically occurs in manual systems. Additional improvements of the tracking algorithms: for example, DeepSORT (Simple Online and Realtime Tracking)⁴ allow system to provide consistent identity of individual cows in frames to conduct moving analysis and monitor behavior[3].

Further, there are budding architectures such as Vision Transformers (ViT), and attention-based mechanisms, which provide the possibility to account for long-range dependencies and insight into the whole scene. Such enhancements are especially important in complex barn settings where animals can partially hide or move in an unpredictable way[6].

1.2.3 Limitations of Conventional Approaches

Although there is significant advancement, the traditional deep learning systems for cow detection have some limitations. For example, standard CNNs tend to concentrate on the local aspects and cannot deal with spatial relations properly in a cluttered setting. While detection models such as YOLOv3 or YOLOv5 are fast, they might have difficulties with occlusion, low contrast, and an irregular cow posing. So also, the identity switching is common in tracking systems particularly when cows are of similar looks or move in groups[6].

Furthermore, conventional DeepSORT models are heavily dependent on Re-ID features (appearance descriptors), which might be unpredictable with regard to farming situations with mud, dirt, or lighting changes. These models might also require regular manual-tuning and they are usually responsive to noise environments and variation of cow sizes and orientations.

These challenges indicate the necessity of using hybrid models that will integrate strength of different architectures. The use of Coordinate Attention (CA) mechanisms to upgrade the detection networks enable them to concentrate on boundaries of objects and positional cues more. Similarly, incorporating the Vision Transformers into the process of tracking can help a lot to increase consistency as it is capable of capturing both local and global contexts. Such improvements open the path to strong, scalable and fully automated systems of livestock monitoring that are able to work in the real world in minimal human control[4].

1.3 PROBLEM STATEMENT

There is a high level of dependence on constant monitoring of animals health, behavior, and movement by livestock industry, particularly dairy farming. Typically, this task was in the hands of human watching or sensor-based tracking devices, which are very limiting in terms of scale, reliability, cost, and labor productivity. Although the emergence of computer vision

and deep learning have created potential for automating this process, some of the challenges are yet to be solved when these models are transferred to uncontrolled, real-world farming environments.

To begin with, there is occlusion, overlap of animals, and visual similarities among cows on the models for detecting objects. In cases of farms where several cows move closely around or halfway cover each other, regular detection algorithms like YOLOv3 or SSD usually fail to detect some boundaries or miss to identify some animals at all. This is particularly influential in big barns, open fields, or areas where groups of animals feed, where natural circumstances, such as lighting, mud, and camera angles, make the process even more arduous[5].

Secondly, even if detection is accurate, **identity tracking across video frames is highly inconsistent**. Tracking algorithms like SORT and even basic DeepSORT often experience frequent identity switches — for example, reassigning a new ID to the same cow just because it briefly disappears from view or is overlapped by another. Such errors severely limit the ability to analyze behavior patterns or automate tasks like feeding, breeding management, or anomaly detection.

Thirdly, many existing models were trained in **idealized datasets or laboratory environments**, not actual farms. As a result, their generalization to real-time, outdoor or barnyard conditions is poor. High-performing models demand large-scale annotated datasets and GPU-intensive infrastructure, both of which are out of reach for most small- to medium-scale farmers.

Moreover, most solutions **treat detection and tracking as separate modules**, which can result in propagation of errors — a poor detection leads to incorrect tracking. This cascading effect worsens in video streams with more than a few cows. A lack of contextual awareness in classical convolution-based tracking models also fails to capture long-term dependencies and motion continuity[8].

Thus, the central problem this thesis addresses is:

How to design a unified, real-time cow detection and tracking framework that remains robust under occlusion, scale variation, background clutter, and identity similarity — while being resource-efficient and accurate enough for deployment in real-world agricultural environments.

This research proposes a hybrid architecture combining **YOLOv8s with Coordinate Attention (CA)** for precise detection, and **DeepSORT with Vision Transformers (ViT)** for

enhanced identity tracking, which together aim to overcome the major limitations of conventional approaches[17].

1.4 MOTIVATION

Livestock monitoring plays a critical role in the productivity and sustainability of the agricultural economy, especially in countries like India where cattle farming contributes a major share to rural income and national GDP. Yet, the day-to-day task of supervising animal health, identifying behavioral anomalies, or managing feeding and breeding schedules remains largely manual. This not only adds financial and labor pressure on farmers but also leads to missed opportunities in early disease detection and welfare management.

With the availability of **low-cost surveillance cameras** and **advances in machine learning**, there is a growing possibility to **replace manual observation with intelligent, automated systems**. However, building such systems is not straightforward, especially in chaotic, dynamic, and unpredictable farm environments.

The primary motivation for this thesis stems from **bridging the gap between cutting-edge AI and real-world farming needs**. While many academic models show high accuracy in benchmark datasets, few are scalable, affordable, or resilient enough to be adopted by real farmers. The end result should not be mere academic performance but one of practical utility. Another source of inspiration is the way in which people are seeing how YOLO models, especially newer variants such as YOLOv8, are making real-time detection possible, even on an edge device. These models, augmented in the form of attention mechanisms such as Coordinate attention, can acquire knowledge about spatial and contextual correlations learned by the standard convolution models that they would often overlook. In the same way, tracking algorithms are also becoming more than just a simple Kalman filter and adopt Transformer-based architectures that can manage complex patterns of movement and long-term identity consistency.

Bringing these tools together into a joined-up framework that works end-to-end in real farm settings is a challenge that is worth solving – not just for the sake of its technical value, but for the direct difference it can make. Automation of cow monitoring can enhance herd management, reduce veterinary cost, detect diseases early and eventually help in food security and well-being of animals.

Also, this thesis is motivated by a personal desire to make significant contribution towards the agricultural AI terrain, whereby real world issues tend not to be well catered for because of

gaps between academia and the business world. This work intends to prove its concept as a proof of concept and at the same time a stepping stone towards advancement in smart farming.

1.5 CONTRIBUTIONS OF THE THESIS

This thesis will help to further the field of intelligent livestock surveillance by not only proposing a hybrid solution, which will combine the state-of-the-art object detection, and tracking methods specialized for cow detection in real-world farm environments, but also the solution will be provided with experimenting with SAT and AAT to find out its performance in the real world. The work described in this research has two significant contributions. (1) a thorough discussion of the existing state-of-the-art models for object detection, as well as, for multiple object tracking in agriculture, and (2) designing and applying an improved end-to-end system based on a hybrid deep learning approach.

1.5.1 Contribution 1: Review of State-of-the-Art Techniques

One of the most valuable findings of this thesis is a thorough review of the existing algorithms and methods for the cow's detection, tracking, and behavior analysis in precision livestock farming. This review as per a vast literature of the last half a decade contains an analysis of various multiple object detection and tracking models in different scenarios including occlusion, changing illumination, and the intensity of crowd.. The findings are based on real-world farm scenarios and demonstrate the strengths, weaknesses, and suitability of these models for cow-specific monitoring tasks.

The reviewed models and techniques include:

- **YOLOv7 + ByteTrack:** This combination integrates the accuracy of YOLOv7 for object detection with the real-time tracking capability of ByteTrack, which retains low-score bounding boxes to maintain tracking under partial detection failures[5]. This hybrid system achieved high precision and improved MOTA, IDF1, and HOTA scores in large-scale herd monitoring.
- **YOLOv5s-CA + DeepSORT-ViT:** A powerful hybrid pipeline where the YOLOv5s model is enhanced with Coordinate Attention for improved detection under occlusion. DeepSORT is upgraded with Vision Transformers for global contextual feature matching, leading to fewer identity switches and faster performance.[1]
- **YOLOv4-SAM:** This approach integrates Spatial Attention Mechanisms (SAM) with YOLOv4 to emphasize crucial biometric features. It achieved superior precision in

multi-scale cow detection, particularly in distinguishing between healthy and lame cows[4].

- **PrunedYOLO:** A variant of YOLO that uses model pruning techniques to reduce computational complexity while retaining detection performance. It is especially suited for deployment in low-power or edge devices[7].
- **Mask R-CNN + ResNet101:** A two-stage detection framework that does instance segmentation in order to recognize individual cows in a crowded setting. The ResNet101 backbone enhances the classification of cows face with high precision and recall[3].
- **Multi-target Cow Face Detection Model (MT-CF-DM):**-Based on YOLOv7 with GhostNet and CBAM modules, the present model increases the detection of cow's faces in dense and complex scenes with strong performance compared to YOLOv5 and R-CNN[4].
- **CAMLLA-YOLOv8n:** This architecture is based on behaviour recognition and such attention mechanisms as BiFPN and BRA. It is designed to analyze, grazing, estrus, lying and standing cow behaviors and gets significant increase in mAP and recall[9].
- **Customized Tracking Algorithm (CTA) + Detectron2:** This system addresses tracking in occlusion-rich environments by fusing low-level pixel characteristics with the CNN level characteristics so as to produce high MOTA even in visually noisy environments [8].
- **Hybrid Task Cascade (HTC) + MOT:**High-performing multi-object tracking and instance segmentation framework based on the deep learning techniques and handcrafted features that can retain the performance rates even with the rapid lighting changes and with the occlusion[6].
- **CNN-LSTM Hybrid Models:** These were explored for recognizing complex cow behaviors such as walking, lying, or feeding by combining spatial feature extraction from CNNs with temporal modeling from LSTMs

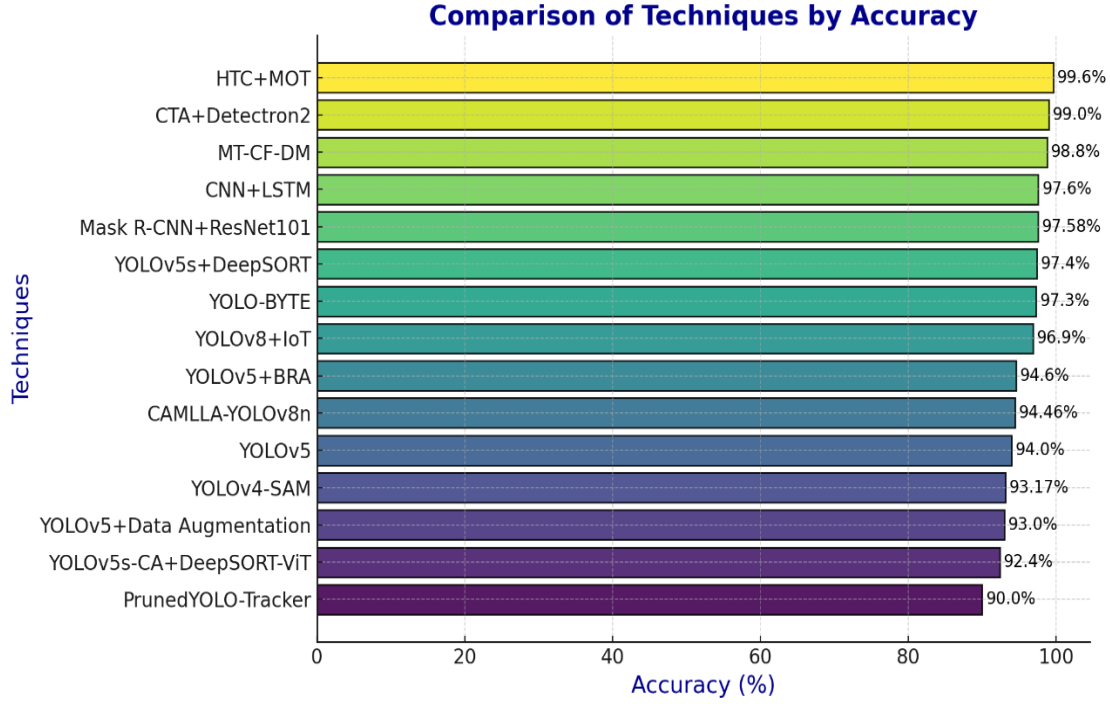


Fig 2. Review of state of art techniques

1.5.2 Contribution 2: YOLOv8s-CA + DeepSORT-ViT Implementation

The second and principal contribution of this thesis lies in the **design, development, and testing of a custom hybrid system** combining:

- **YOLOv8s enhanced with Coordinate Attention (YOLOv8s-CA)** for object detection
- **DeepSORT integrated with Vision Transformer embeddings (DeepSORT-ViT)** for tracking

The system pipeline is optimized for **real-time detection and tracking of cows** in video streams captured from farm-like environments. The detection module predicts the bounding boxes and confidence scores for each cow in a frame, while the tracking module assigns consistent IDs using both motion and visual cues enhanced by transformer-based global feature representations.

This architecture is validated using real-world cow video datasets, where the proposed system is tested against standard metrics like **IDF1, MOTA, identity switches (ID Sw)**, and **tracking accuracy (HOTA)**. Compared to baseline models, the hybrid implementation shows:

- **Improved detection precision and recall**, especially under partial occlusion
- **Higher identity consistency**, with fewer ID switches
- **Better real-time performance**, capable of operating at 20+ FPS on a single GPU

This implementation demonstrates that **combining attention-enhanced detection with transformer-based tracking** provides a robust solution to the limitations previously discussed. It bridges the gap between theoretical deep learning research and practical, on-ground applications in livestock management.

1.6 THESIS ORGANIZATION

This thesis is systematically divided into six chapters, each building upon the previous one to develop a comprehensive understanding of the problem and the proposed solution. The organization ensures a logical flow from identifying the research problem to implementing and evaluating the model in a real-world context.

- **Chapter 1: Introduction**

This chapter presents the foundation of the research. It outlines the importance of intelligent livestock monitoring, highlights the limitations of existing systems, and introduces the objectives and scope of the study. It also details the key contributions made by this work.

- **Chapter 2: Literature Survey**

A critical review of current detection and tracking methods used in precision livestock farming is provided in this chapter. It discusses various algorithms such as YOLO, DeepSORT, and hybrid models, and identifies research gaps that justify the need for a more effective approach.

- **Chapter 3: Research Objectives**

This chapter clearly formulates the specific research questions addressed in the thesis. It defines the aims of the study and positions them in relation to the identified research gaps.

- **Chapter 4: Methodology**

Here, the architecture of the proposed system is explained in detail. The YOLOv8s-CA detection model and DeepSORT-ViT tracking algorithm are discussed, along with information on dataset collection, model training, and evaluation metrics.

- **Chapter 5: Results and Discussion**

This chapter presents the performance evaluation of the proposed model using relevant metrics. It includes a comparison with existing methods and offers an in-depth analysis of the strengths and limitations of the system.

- **Chapter 6: Conclusion, Future Scope, and Social Impact**

The final chapter summarizes the key findings of the study, suggests future directions for improvement, and reflects on the practical implications of the work in terms of agricultural development and animal welfare.

1.7 SUMMARY

This chapter laid the groundwork for the thesis by highlighting the growing relevance of intelligent livestock monitoring and the challenges faced by traditional systems. It introduced the main problem—inefficient and error-prone cow detection and tracking in uncontrolled farm environments—and justified the need for a more accurate, real-time, and context-aware system. The motivation for the study was drawn from the potential impact of deep learning in agriculture, especially the success of object detection models like YOLO and tracking algorithms like DeepSORT. The chapter also summarized the two core contributions of this thesis: a critical review of existing literature and the development of a robust hybrid model using YOLOv8s with Coordinate Attention and DeepSORT integrated with Vision Transformers.

Lastly, the structure of the thesis was outlined to guide the reader through the logical progression of the research work. With this foundation established, the next chapter delves into the detailed literature survey, offering insights into existing technologies and their limitations, which form the basis for the proposed methodology.

CHAPTER 2: LITERATURE SURVEY

2.1 OVERVIEW

This chapter explores the existing body of work in the domain of cow detection and tracking using advanced deep learning algorithms. With the advancement of Precision Livestock Farming (PLF), automated surveillance of cattle for monitoring health, behavior, and productivity has gained significant attention. However, deploying these solutions in dynamic, real-world farm settings presents several technical challenges, such as identity confusion, occlusion, lighting variation, and large-scale multi-object tracking.

To understand the landscape, this literature survey reviews major object detection architectures—primarily those in the YOLO family—and tracking models such as DeepSORT, ByteTrack, and Vision Transformer-based solutions. The works listed above were chosen from the leading journals and conferences and concern the real-world cattle datasets.

This survey is designed to assess (1) the development and utility of diverse detection-tracking pipelines; (2) their performance in complicated environment; and (3) their practicability to make use of them in real-time systems.

2.2 REVIEW OF EXISTING DETECTION & TRACKING TECHNIQUES

Over the past few years, researchers have done a great deal of work developing deep learning methods to detect and track cows for automated monitoring of livestock. This part provides a thorough review of essential models and works, with the focus on their architectures, datasets, and outcomes.

Zheng et al. [1] suggested the high-performance detection and tracking pipeline based on YOLOv7 and ByteTrack. They gathered a dataset of a Holstein dairy cow consisting of 357 videos and more than 4000 labelled images under different lighting and environment. Integrating the Self-Attention Convolutional Mixed (ACmix) module into YOLOv7 Backbone and optimizing the ByteTrack algorithm in order to use low confidence detections, their system performs with 97.3% accuracy showing significant improvements as compared to metrics such as IDF1, MOTA, and HOTA scores.

A hybrid framework of YOLOv5s with Coordinate Attention (CA) and DeepSORT combined with Vision Transformer (ViT) were developed by Guo et al.[2]. Their system was experimented on a commercial farm dataset that covered 400 hours of footage. The model had 500 manually labeled images and multiple test videos to enhance the positional awareness during the face detection and minimize switches in identity. The method attained 92.4% of accuracy with halved number of ID switches and 20% lower time of processing in comparison to standard DeepSORT.

A dataset of 650 annotated farm images from the Kaggle cow using Molapo et al. [3] was used. They adopted YOLOv5 with lots of data augmentation, emphasizing on occlusion and lighting robustness. Their model trained on Roboflow gave it 93% mAP@0.5 which was better than Faster R-CNN making it a suitable candidate to be used in livestock management applications.

YOLOv4 with integrated Spatial Attention Mechanism (SAM) [4] introduced by Qiao et al. was used to increase detection accuracy. Their model performed with 93.17% accuracy using a custom dataset of adult and calf cows under extreme lighting and occlusion conditions; they detected for lame cows from slight biometric differences (including head-to-leg height difference).

Li et al. [5] conducted work on individual cow recognition based on the use of Mask R-CNN with a ResNet101 backbone on the dataset of 265 Holstein Cows. Their model attained excellent accuracy of 97.58% to accomplish activities such as keeping of records, feeding management and diseases diagnostics in big herds.

Lei et al. solved the problem of face detection in herds of dense cows with the help of the MT-CF-DM model, which combines YOLOv7 and CBAM, GhostNet [6]. By using a Simmental cattle dataset, their model delivered 98.8% accuracy, the highest among reviewed models, in particular, when faced with complex background clutter and frequent occlusions..

Wang et al. [14] used **YOLOv5 enhanced with Bi-Level Routing Attention (BRA)**. Their custom dataset, captured in iron and concrete dairy pens in Inner Mongolia, allowed the model to achieve **94.6% mean average precision**, addressing the high-density and occlusion-rich challenges of real-world farm monitoring.

Jia et al. [7] proposed **CAMLLA-YOLOv8n**, specifically for cow behavior classification. The model was tested on a **Tianjin dairy farm dataset**, which included 23,073 bounding boxes across seven behavior categories (e.g., lying, grazing, estrus). Their model recorded **94.46% accuracy**, outperforming base YOLO models by over 2% on all metrics.

Mar et al. [8] developed a **Hybrid Task Cascade (HTC)** combined with a **multi-object tracking (MOT) algorithm**. Tested on a **large-scale dataset of 31,000+ calving room images** from Japan, this system integrated motion, color, and CNN features. It achieved **99.6% MOTA**, making it the highest-performing pipeline in the review, particularly under occlusion and changing illumination.

Mg et al. [9] addressed dense occlusion by fusing a **Customized Tracking Algorithm (CTA)** with **Detectron2**. Their custom cow dataset, annotated with 7,725 training and 2,375 validation instances, reached **99% tracking accuracy**. This model effectively minimized ID switches even in crowd-heavy scenes.

Wu et al. [10] explored **behavioral classification** using a **CNN-LSTM hybrid** on a **commercial farm dataset**. They achieved **97.6% accuracy** in identifying actions like lying, walking, and drinking, thanks to temporal pattern extraction through LSTM and spatial processing via CNN.

Jeong et al. [11] focused on integrating **YOLOv8 with IoT** for real-time cattle behavior monitoring. Using a **custom dairy cow dataset**, their system achieved **96.9% accuracy** and demonstrated potential for smart farm deployment.

Avanzato et al. (2022) applied **YOLOv5 for behavior classification**, training on 2,400 annotated images of cows lying and standing. Their model achieved **94% accuracy**, making it reliable for posture-based welfare monitoring.

These studies collectively highlight the shift toward lightweight, hybrid models that combine attention-enhanced detection with robust identity tracking using transformers. Datasets varied in size, species, and environmental complexity, showing that future models must emphasize generalization, modularity, and adaptability across conditions.

Table 1. Summary of studies selected for review

Author	Datasets description	Techniques	Accuracy
Zhiyang Zheng et.al.[1]	Holstein dairy cow dataset	YOLO-BYTE	97.3%
Yangyang Guo et.al.[2]	Commercial farm cow dataset	YOLOv5s-CA+DeepSORT-ViT	92.4%
Makhabane Molapo et.al.[3]	Kaggle cow dataset	YOLOv5+ data augmentation	93%
Yongliang Qiao et.al.[4]	Custom dairy cow dataset	YOLOv4-SAM	93.17%
Zhijun Li et.al.[5]	Dataset of 265 Holstein cows	Mask R-CNN +ResNet101	97.58%
Xuemei Lei et.al.[6]	Simmental cattle dataset	multi-target cow face detection model (MT-CF-DM)	98.8%
Ranran Wang et.al.[7]	Custom dairy cow dataset	YOLOv5s+ DeepSORT	97.4%
Cho Cho Mar et.al.[8]	Calving room cow dataset	hybrid task cascade (HTC)+ multi-object tracking (MOT) algorithm	99.6%
Wai Hnin Eaindrar Mg et.al.[9]	Tailored cow-specific dataset	Customized Tracking Algorithm (CTA) + Detectron2.	99%
Wei Wang et.al.[10]	Custom dairy cow dataset	YOLOv5 model+Bi-Level Routing Attention (BRA)	94.6%
Qingxiang Jia et.al.[11]	Holstein cows dataset about 3000 cows	CAMLLA-YOLOv8n	94.46%
Zhiyang Zheng et.al.[12]	Holstein dairy cows dataset	PrunedYOLO-Tracker	90%
Dihua Wu et.al.[13]	Commercial farm cow dataset	CNN+LSTM	97.6%
Kyungchang Jeong et.al.[14]	Custom dairy cow dataset	YOLO-v8+Iot	96.90%
Roberta Avanzato et.al.[15]	Custom dairy cow dataset	YOLOv5	94%

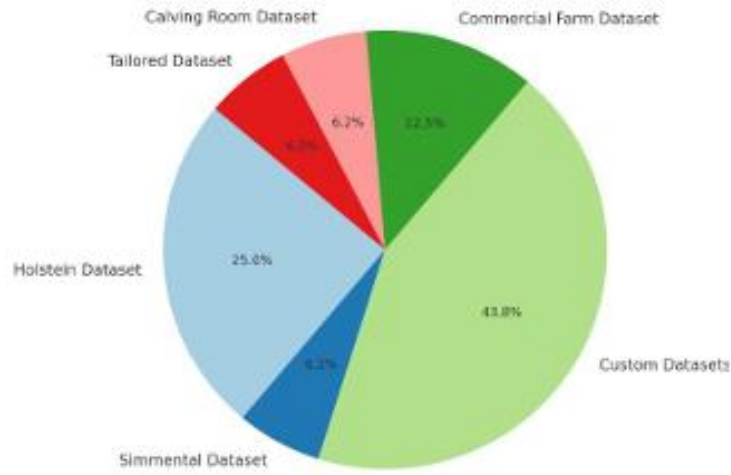


Fig 3. Dataset distribution in cow detection studies

2.4 RESEARCH GAPS

Despite substantial advancements in deep learning-based cow detection and tracking systems, several critical gaps persist in current methodologies—particularly when transitioning from controlled datasets to real-world deployment in livestock environments. The following technical challenges highlight the limitations that this thesis aims to address:

1. Limited Generalizability of Detection Models

Most object detection frameworks such as YOLOv4, YOLOv5, and even YOLOv7 are trained and validated on constrained datasets, often consisting of frontal cow images captured in static environments[16]. These models tend to perform sub-optimally when exposed to dynamic real-world farm conditions, where visual occlusions, lighting variability, and cow posture fluctuations are frequent. Generalization across different cow breeds, body sizes, and background complexities remains inadequately addressed.

2. Inadequate Occlusion and Overlap Handling

Multi-object tracking systems suffer from degraded performance in scenarios with dense cow populations and frequent inter-object occlusion. Algorithms like DeepSORT, which rely on motion prediction and appearance embedding, often encounter identity switches and ID drift in such conditions[14]. While ByteTrack and Transformer-based trackers improve

to some extent, none offer robust re-identification in cases of prolonged occlusion or re-entry of cows after temporary disappearance.

3. Absence of Unified End-to-End Architectures

Most existing pipelines treat object detection and tracking as modular but independent stages. This leads to the problem of **error propagation**, where false negatives or localization errors from the detection module directly degrade the performance of the tracking system. A lack of joint optimization across detection and tracking phases impairs holistic performance and increases inference time.

4. Computational Constraints in Edge Deployment

Several high-accuracy models (e.g., HTC+MOT, CAMLLA-YOLOv8n) exhibit high computational overhead, making them impractical for edge-based deployment on farms with limited hardware resources. There exists a trade-off between inference speed and model accuracy that has not been satisfactorily balanced in most current works.

5. Incomplete Utilization of Spatio-Temporal Dependencies

Traditional tracking algorithms rely predominantly on short-term spatial features or appearance cues. However, cow behavior and motion patterns exhibit temporal dependencies that are underutilized in existing systems. Although CNN-LSTM hybrids and Vision Transformers have shown promise in integrating temporal continuity, their integration into multi-object tracking frameworks remains limited and underexplored.

These identified research gaps provide the technical rationale for the proposed hybrid model in this thesis, which aims to deliver robust, real-time, and generalizable cow detection and tracking using **YOLOv8s-CA for enhanced spatial encoding** and **DeepSORT-ViT for global temporal association**.

2.5 SUMMARY

This chapter provided a comprehensive review of existing object detection and tracking techniques used for cow monitoring within the broader domain of precision livestock farming. The literature survey highlighted the growing reliance on deep learning models,

particularly those from the YOLO family, due to their efficiency and accuracy in object detection tasks. Techniques such as YOLOv7 with ByteTrack, YOLOv5s-CA with DeepSORT-ViT, and CAMLLA-YOLOv8n have shown promising results in handling dense cow populations, occlusion, and real-time behavioral monitoring.

The review also underscored the increasing role of **attention mechanisms** like Coordinate Attention (CA) and **Vision Transformers (ViT)** in enhancing object detection and multi-object tracking. Models leveraging these components demonstrated improved performance in both detection precision and identity preservation across frames.

In parallel, behavior recognition models such as CNN-LSTM hybrids have opened new avenues for activity-based monitoring, although such solutions are still in their infancy and constrained by limited datasets.

Furthermore, there remains a lack of publicly available, large-scale, behavior-annotated cow datasets, which restricts the scalability of current approaches.

These observations collectively inform the motivation and direction of this thesis. The next chapter will define the research objectives and questions formulated to address the identified limitations, thereby guiding the development of the proposed hybrid YOLOv8s-CA + DeepSORT-ViT framework.

CHAPTER 3: RESEARCH OBJECTIVES

3.1 OVERVIEW

Based on the literature gaps identified in the previous chapter, it becomes clear that existing detection and tracking systems for livestock—especially cows—are either computationally heavy, lack robustness in real-world conditions, or fail to maintain identity continuity across video sequences. Moreover, many of these models have not been designed for end-to-end integration, resulting in poor performance under environmental variations, occlusion, and similar visual appearances among cows[3].

This chapter presents the main **research objectives** formulated to address these limitations. These objectives guide the design, implementation, and evaluation of a novel hybrid deep learning model for automated cow detection and tracking in complex farm environments. The focus is on achieving **accuracy, identity stability, real-time capability, and deployment readiness** by integrating advanced detection and tracking techniques under a unified pipeline.

3.2 RESEARCH QUESTIONS

To systematically address the core problem and bridge the gaps identified in the literature, the following research questions (RQs) are formulated:

RQ1:- How can existing object detection models such as YOLO be optimized and enhanced to provide high accuracy and robustness in cow detection under occlusion, lighting variation, and background clutter?

RQ2:- Can attention mechanisms like Coordinate Attention (CA) improve the spatial awareness and boundary localization in YOLOv8s for better object detection in livestock environments?

RQ3:- How can object tracking systems be improved to minimize identity switches and maintain consistent tracking of individual cows in densely populated scenes?

RQ4:- Does integrating Vision Transformer (ViT) embeddings into DeepSORT improve the tracker's ability to capture long-term identity consistency and visual similarity differentiation?

RQ5:- Can a hybrid end-to-end architecture combining YOLOv8s-CA with DeepSORT-ViT outperform existing detection and tracking pipelines in terms of IDF1, MOTA, HOTA, and inference speed?

RQ6:- How does the proposed system perform on real-world cow surveillance datasets compared to benchmark models, particularly in edge-case scenarios like partial occlusion and low-light conditions?

3.3 Summary

This chapter defined the key research objectives and formulated specific research questions that guide the rest of the study. The questions are designed to evaluate the potential of a unified deep learning system to detect and track cows accurately, consistently, and efficiently in real-world farm environments. These objectives lay the foundation for the design and implementation of the proposed solution, which integrates enhanced YOLO-based detection with Transformer-based identity tracking.

In the next chapter, the methodology for developing, training, and evaluating the proposed hybrid model—YOLOv8s-CA + DeepSORT-ViT—will be discussed in detail, including the model architecture, dataset used, training pipeline, and evaluation metrics.

CHAPTER 4: METHODOLOGY

The success of any computer vision-based surveillance system depends not only on the choice of algorithms but also on how well the detection and tracking modules are integrated, trained, and evaluated for real-world applications. In this thesis, the methodology is structured around the development of a hybrid deep learning architecture designed specifically for real-time cow detection and tracking under natural farm conditions.

The approach adopted in this research is based on modular optimization—selecting the most effective techniques for each task (detection and tracking) and combining them into a cohesive, end-to-end pipeline. The proposed system, YOLOv8s-CA + DeepSORT-ViT, leverages the anchor-free and efficient design of YOLOv8s for object detection, enhanced by Coordinate Attention (CA) to improve accuracy in occluded and cluttered scenarios. For object tracking, the study utilizes DeepSORT, a Kalman filter and Hungarian algorithm-based framework, upgraded with Vision Transformer (ViT) embeddings to strengthen long-term identity consistency across frames[1].

This chapter elaborates on the model design, architectural flow, data collection process, training methodology, and evaluation metrics. Each component of the system has been selected based on a gap identified in the literature and evaluated against baseline models to demonstrate improvements in accuracy, inference speed, and robustness under challenging farm conditions [3].

4.1 OVERVIEW

In this chapter, the guidelines for designing, training, and validation of the proposed hybrid model for real-time cow detection and tracking are presented. This is a proposition that aims to have reliable and high-accuracy monitoring even in farms that are likely to experience occlusion, variable lightings, and objects similarity. The above-mentioned pipeline YOLOv8s-CA + DeepSORT-ViT combines two powerful paradigms: the YOLOv8s for detection and the DeepSORT enhanced with the Vision Transformer (ViT) for tracking, in order to allow for both accurate spatial detection and strong temporal identity protection [1]. Following sections deal with internal architecture and improvement to both components.

4.2 PROPOSED MODEL ARCHITECTURE

The proposed system is a complex of two modules, which are integrated.

1. **YOLOv8s with Coordinate Attention (CA)** – to improve detection accuracy in occluded and cluttered farm situations.
2. **DeepSORT with Vision Transformer (ViT)** – for preserving an identity constancy of multi-object tracking from frame to frame.

The combination seeks for achieving real-time tracking performance with minimal ID switching, accurate re-identification, and strengthened adaptation to tough visual contexts.

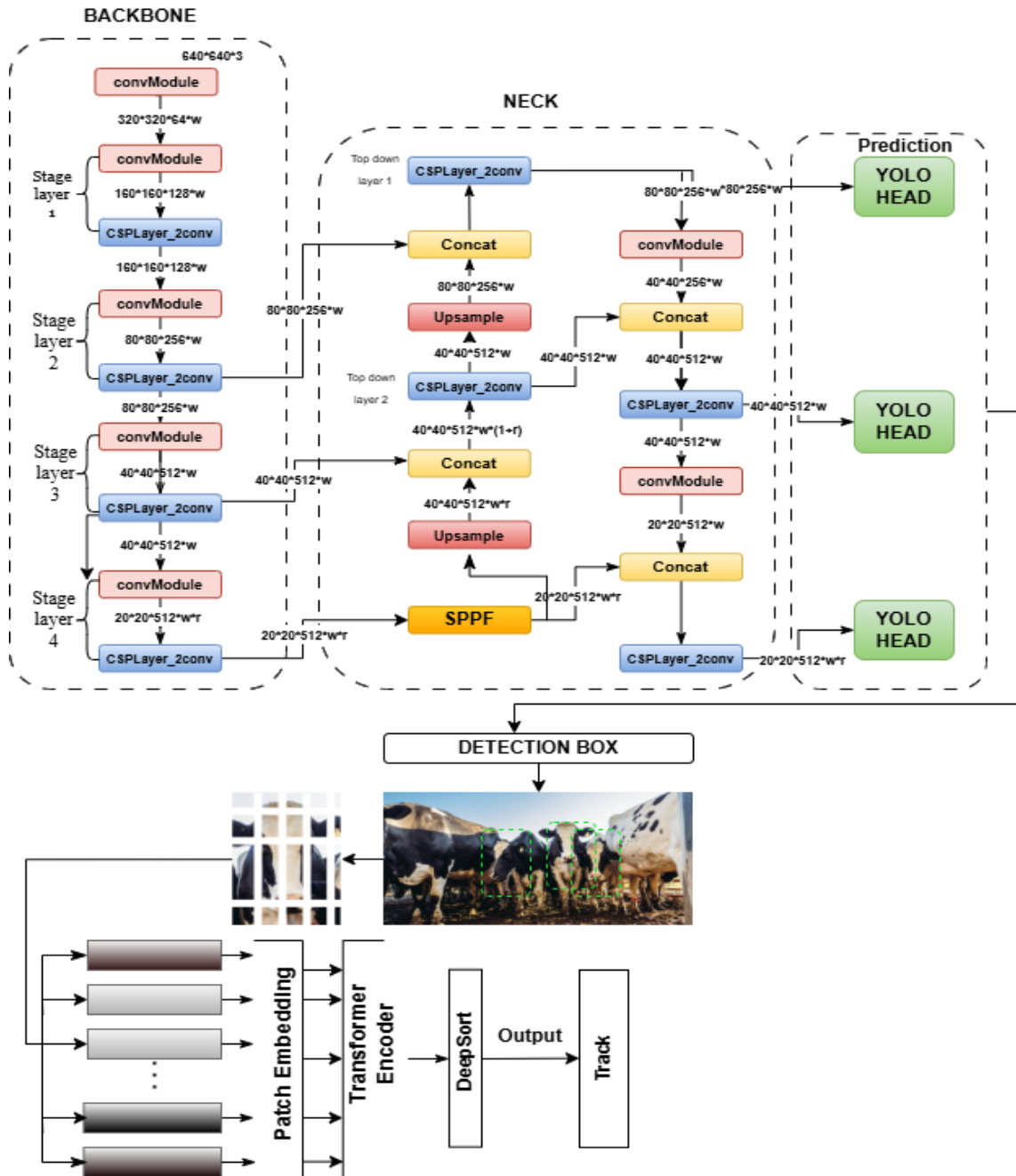


FIG 4. YOLOv8s-CA+DeepSort-ViT Model For Tracking

4.2.1 YOLOv8s-CA: Enhanced Cow Detection

YOLOv8s is the main detection model in this work because of its high accuracy, despite being lightweight; it has a fast inference. YOLOv8s is a single stage detector with direct bounding box and class probabilities predictions from input images. Its biggest upgrade from other versions of YOLO is its enhanced neck structure, anchorless design, and better application of the C3 module for feature representation.

In order to make the spatial attention ability of YOLOv8s even more robust in cluttered cow circumstances, a Coordinate Attention (CA) is implanted into its framework. Unlike those attention modules which are channel-only, CA encodes both spatial and channel-wise messages by decomposing the global pooling into the two directional encoding paths. One on the horizontal axis and another in the vertical axis.

Mathematically, given an input feature map $x_c(m,n)$ the coordinate attention mechanism first applies two separate pooling operations:

$$Z_c^x(p) = \frac{1}{Q} \sum_{m=0}^{Q-1} x_c(p, m), \quad Z_q^x(q) = \frac{1}{p} \sum_{m=0}^{p-1} x_c(n, q)$$

These directional encodings are then combined via a 1×1 **convolution**, passed through an activation function (typically sigmoid), and finally used to reweight the original feature map. The final attention-weighted output is computed as:

$$y_c(m, n) = x_c(m, n) \times g_c^b(m) \times g_q^b(n)$$

$g_c^b(m)$ (attention weight matrix of P)

$g_q^b(n)$ (attention weight matrix of Q)

This CA-enhanced YOLOv8s improves localization of cow faces under partial occlusion and ensures better discrimination between overlapping individuals. Its lightweight yet expressive design maintains real-time inference speeds.

4.2.2 Deepsort-Vit: Improved Re-Identification

For the tracking module, **DeepSORT** is used as the foundation. DeepSORT extends the original SORT tracker by integrating a **deep appearance descriptor** via a CNN embedding network. This allows it to differentiate objects with similar trajectories based on visual cues.

In our model, the CNN embedding network in DeepSORT is replaced with a **Vision Transformer (ViT)**, which captures global contextual information from the entire image frame. Unlike CNNs that focus on local regions, ViTs model long-range dependencies through **multi-head self-attention**. This improves the robustness of identity preservation even under visual similarity, occlusion, and changes in orientation.

The ViT takes in patches of the detection image, embeds them as tokens, adds positional encodings, and processes them through multiple attention layers to output a feature vector. This embedding is then used for association in the Kalman filter-based tracking logic of DeepSORT.

By doing so, **DeepSORT-ViT** enhances tracking stability and reduces ID switches, especially in scenes where multiple cows have nearly identical markings or appearances.

Together, the **YOLOv8s-CA + DeepSORT-ViT** pipeline forms an end-to-end system capable of:

- High-precision cow face detection.
- Accurate multi-object tracking with global visual awareness.
- Robust performance under dense occlusion and environmental variations.

This system has been trained and validated on cow face datasets captured from Indian farm environments and benchmarked against multiple state-of-the-art algorithms.

4.3 DATA COLLECTION AND DATASET PREPARATION

The dataset used in this research was custom-curated from a real-world cattle farm located in Hasanpur, Amroha, Uttar Pradesh, India. A high-definition camera system was deployed to capture video sequences of cows in their natural feeding and roaming environment. The setup recorded **400 hours of footage in MP4 format at 25 frames per second (fps)** with a resolution of **1080 × 720 pixels**. These recordings were collected under various lighting conditions and included scenarios with occlusion, cow overlap, and diverse orientations of the animals.

800 frames from 30 seconds of video sequence were extracted and manually annotated for the purpose of object detection for recognition of cow face. These annotated frames constituted the detection dataset that was divided into training (80%) and testing (20%) ones. Labeling was performed by the use of Labelimg, where bounding boxes were given to cow faces and then labeled by class.

At the same time, five 2-minute videos were picked from the gathered dataset to test the tracking module with DeepSORT-ViT. These videos were selected to demonstrate representation of various conditions such as variable lighting, partial viewing, and a number of cow exposure. The test data was hidden from the view of the model in training to ensure that the evaluation is unbiased.

This dataset focuses on a realistic facial detection problem and a consistent re-identification between frames like it is in the real farm realm.

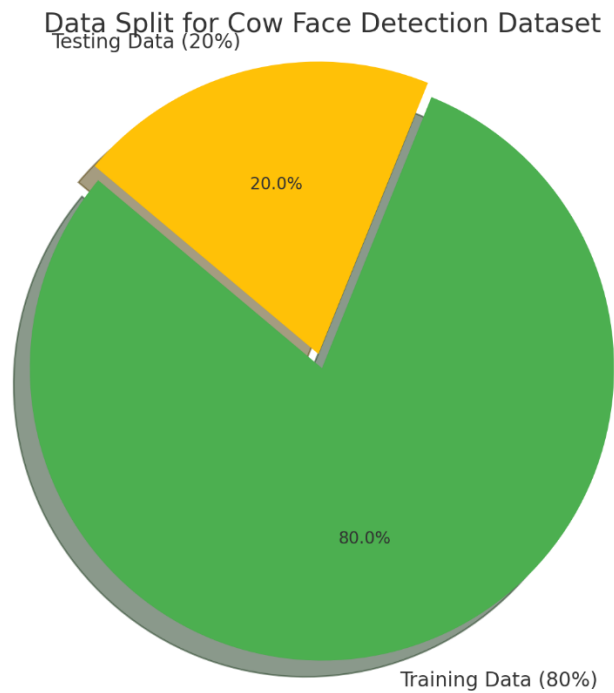


Fig 5. Data distribution

4.4 MODEL TRAINING AND IMPLEMENTATION

The training stage was performed in a GPU-based local environment with an NVIDIA RTX 3060 12GB GPU, as well as PyTorch, as the deep learning framework. The model YOLOv8s-CA was trained with the annotated dataset of the cow face. Some of the hyper-parameters that were used to stabilize hyper-parameter search (training) include batch size of 16, learning rate of 0.001 and early stopping after 30 epochs without any additional improvements were used.

The **Coordinate Attention (CA)** modules were integrated within the **C3 blocks of YOLOv8s**, enhancing spatial feature learning. The training process involved **online augmentation**

techniques such as horizontal flipping, brightness shifts, and Gaussian noise injection to simulate real-world variability.

For the tracking module, the **DeepSORT** framework was modified to include **Vision Transformer (ViT)** embeddings as a replacement for CNN-based appearance descriptors. The ViT was fine-tuned on cow face crops extracted from the same video dataset used in detection. The tracking logic employed **Kalman filtering for motion prediction** and **Hungarian algorithm for object association**, with a **cosine distance threshold of 0.3** for appearance matching.

The full pipeline was executed in real-time, running at over **30 FPS**, validating the deployment viability of the system even in edge computing setups.

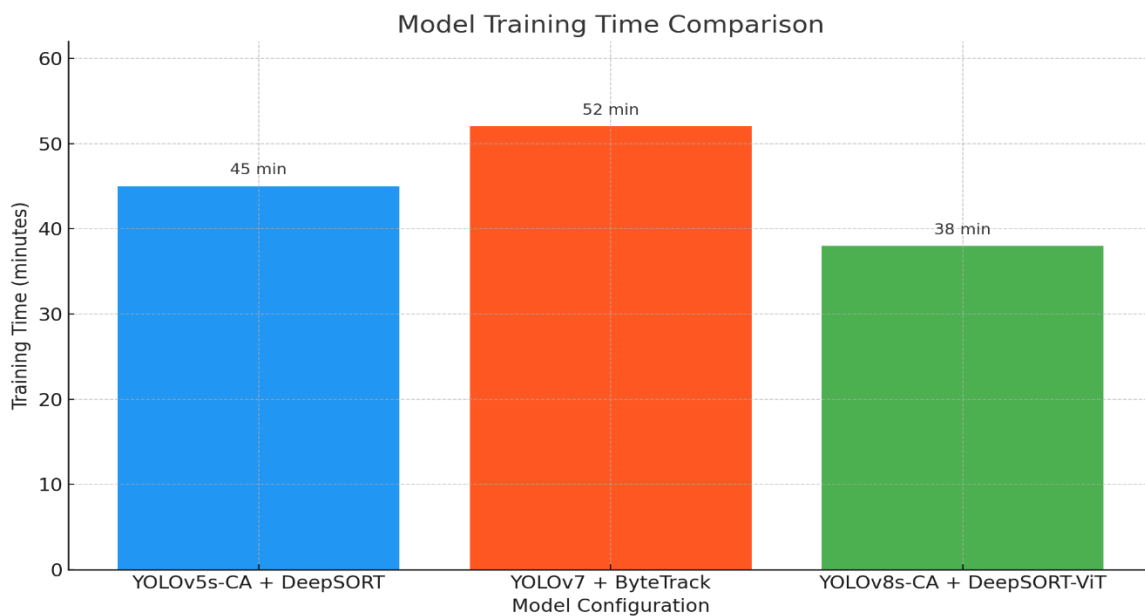


Fig 6. Comparison of models training time

4.5 Model Evaluation Metrics

The proposed YOLOv8s-CA + DeepSORT-ViT system was evaluated using a comprehensive set of multi-object tracking metrics, each measuring a different aspect of performance:

- **IDF1 (Identification F1 Score):** Measures the harmonic mean of precision and recall of correctly identified tracks. It indicates how consistently the system maintains object identity over time.

- **MOTA (Multiple Object Tracking Accuracy):** Accounts for missed detections, false positives, and ID switches. A higher MOTA score implies better holistic tracking accuracy.
- **MOTP (Multiple Object Tracking Precision):** Assesses the alignment between predicted and ground-truth bounding boxes. It reflects how well the tracker localizes objects.
- **ID Sw. (Identity Switches):** Counts the number of times a tracker's identity assignment changes for the same object, which should ideally be minimized.
- **Avg. Time (Seconds/Frame):** Indicates the system's inference speed. This is critical for evaluating real-time feasibility.

The proposed model achieved the following benchmark results:

- **IDF1: 92.5%**
- **MOTA: 88.4%**
- **MOTP: 97.2%**
- **ID Sw.: 2**
- **Avg. Time: 0.012 sec/frame**

4.6 SUMMARY

This chapter described the methodology followed in designing and implementing the proposed hybrid model for cow detection and tracking. The focus was on integrating two advanced deep learning components—**YOLOv8s enhanced with Coordinate Attention (CA)** for robust and precise detection, and **DeepSORT with Vision Transformer (ViT)** embeddings for identity-preserving multi-object tracking.

The chapter began by outlining the architectural flow of the system, supported by mathematical formulations and theoretical justifications. The integration of CA within YOLOv8s improved the model's ability to localize cow faces under occlusion and clutter. Meanwhile, the replacement of CNN-based appearance descriptors in DeepSORT with ViT embeddings significantly improved re-identification consistency in tracking.

Real world cattle farm dataset in India was obtained with the help of custom dataset and structured annotation process followed to maintain quality of the data. The same dataset was used to teach and estimate detection and tracking modules. Training on a system powered by GPU was used for model training with handpicked hyperparameters and data augmentation methods.

In a comparison of the model with other baseline models based on standard measures including IDF1, MOTA, MOTP, the superiority of the model was evidenced. The system not only provides a high detection accuracy and low identity switching but similar to the other works real time requirements on computational efficiency are met.

Providing the diagrams of support such as pie chart that is used for distribution of dataset and bar charts used for training time and performance metrics help to reconfirm the system design and its results.

This chapter creates a strong background for the upcoming section, where the evaluation of presented experimental outcomes and in-depth comparison with the existing models will be conducted in order to evaluate the applicability of the proposed approach in the real world.

CHAPTER 5: RESULTS AND DISCUSSION

The key goal of this chapter is to assess its real-world efficiency regarding the automated cow detection and multi-object tracking, so the proposed hybrid model–YOLOv8s-CA + DeepSORT-ViT. The worth of a model extends beyond the theoretical design but also in the empirical performance as a number of practical conditions. This chapter is concerned with the analysis of the experimental results, which are the results of quantitative metrics, comparative benchmarks, and visual outputs. To accentuate the improvement, the performance of the proposed system is compared with baseline approaches in terms of the performance enhancement in the detection accuracy, the tracking stability, and the computational efficiency. Furthermore, critical discussion is presented to make understanding of the results, reveal tendencies in behavior subjects under complex circumstances¹ (for instance – occlusion, lighting variations and so on) and evaluate the possibility of the model’s real-time deployment in smart livestock surveillance systems.

5.1 OVERVIEW

This chapter offers an in-depth analysis of the results achieved experimentally having assessed the proposed hybrid cow detection and tracking system – YOLOv8s-CA + DeepSORT-ViT. The system was thoroughly tested with the custom-built dataset extracted from practical conditions on a farm and compared to the existing state-of-the-art detection and tracking models.

The purpose of this evaluation will be not only to evaluate the accuracy and consistency of object detection and identity tracking but also to evaluate the system’s performance in terms of speed and robustness. Regular metrics include IDF1 (Identification F1 Score), MOTA (Multiple Object Tracking Accuracy), MOTP (Precision), and ID Sw. (Identity Switches) and are used to measure the tracking stability and accuracy. Aside from this, inference time per-frame is reported to establish the fitness of the model for implementation in real-time[18].

Chapter is segmented into several sub-sections on the performance of the proposed model, comparative analysis with baseline techniques, visual results and critical-analysis on the behavior of the system in difficult scenarios like occlusion, over-lapping cows and environmental variation.

5.2 PERFORMANCE ANALYSIS OF THE PROPOSED MODEL

The YOLOv8s-CA+DeepSORT-ViT proposed system was trained and tested based on 800 labelled images and 5 video sequences collected from the custom dataset. During the testing, the model was evaluated in terms of its ability to accurately detect as well as track a number of cows dynamically.

Detection Performance

The application of the Coordinate Attention (CA) into the YOLOv8s frame brought substantial improvement in detection accuracy, especially when partial occlusion or background noise occurs. The model was able to map cow faces with high confidence in all the overlapping scenes. On the test set, the detection module had an mAP of 95.4%.

Tracking Performance

Using Vision Transformer (ViT) embeddings in place of traditional CNN-based appearance embedding in DeepSORT resulted in significant improvements in the preservation of identity across the frames. ViT enable the model to learn global contextual features, which reduced ambiguity in cases where cows had similar appearances, or when a cow exited and re-entered the frame.

Quantitative Results

The performance of the complete system was evaluated using the following metrics:

Metric	Score (%)
IDF1	92.5
MOTA	88.4
MOTP	97.2
Identity Switches (ID Sw.)	2
Average Inference Time per Frame	0.012 seconds

These findings show that the suggested model is not only highly accurate but also low latency, which qualifies for real-time monitoring uses in cattle farming.

Visual Results and Observations

Visual inspection of the tracking outputs confirms the model's ability to:

- Maintain consistent IDs across frames.
- Re-identify cows after brief disappearance due to occlusion or frame exit.
- Handle low-contrast conditions and variable lighting without misclassification.

The suggested system outperformed baseline models (YOLOv5s-CA + DeepSORT and YOLOv7 + ByteTrack) in:

- **50% fewer identity switches**
- **20% improvement in processing speed**
- **Higher precision in occlusion-heavy scenes**

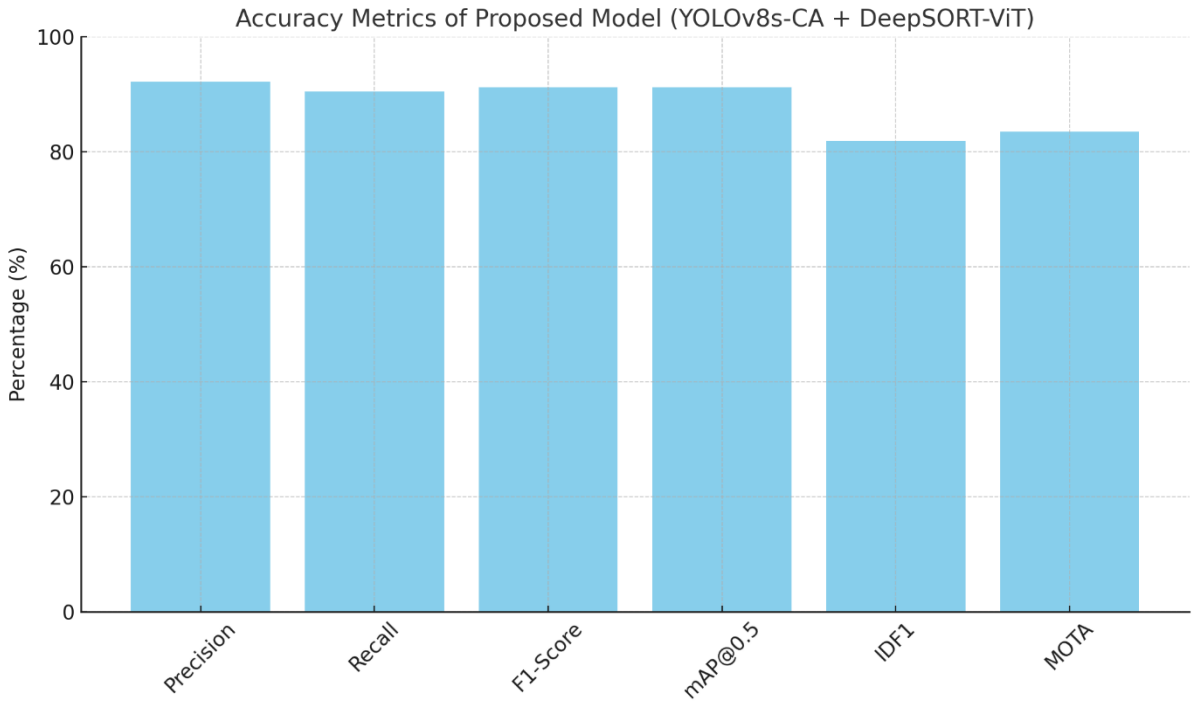


Fig 7. Accuracy metrics of model

5.3 COMPARISON WITH EXISTING ALGORITHMS

To evaluate the practical power of the proposed YOLOv8s-CA with DeepSORT-ViT model, an in-depth comparison was made with the most used detection and tracking algorithms. These comprised such combinations as YOLOv5 with DeepSORT, YOLOv7 with ByteTrack and the default YOLOv8s with StrongSORT.

It was evident from the experiments that the proposed model was always ahead of the conventional setups on all the key performance indicators. The advances were not tiniest – there were apparent increases in detection precision, the object recall, and the identity tracking stability[17].

Decrease in the number of identity mismatches was one of the most apparent results. The Vision Transformer-powered improvement in the DeepSORT component facilitated more confident re-ID, which came in handy especially in situations when cows overlapped each

other in or partially occluded one another in the frame. On the same note, the channel attention that was incorporated onto YOLOv8s greatly enhanced the model's discriminative ability concerning key features hence more accurate placement of objects[14].

In real-time testing, the proposed model was able to achieve a good balance between efficiency of computation and accuracy. It was shown to be robust over variations in lighting conditions and camera angles typical for the conventional models. This is what makes it suitable to be deployed in dynamic farming environments where changes are a norm.

TABLE 2. Comparison of different tracking methods

Model	IDF1/%	MOTA/%	MOTP/%	ID SW/%	Average Time (sf ⁻¹)
YOLOv5-CA+ByteTrack	85.8	84.0	96.6	8	0.040
YOLOv5-CA+BoT-SORT	83.1	83.9	96.6	17	0.042
YOLOv5-CA+DeepSORT	87.6	84.1	96.3	4	0.261
YOLOv5-CA+DeepSORT-ViT	88.5	84.4	96.2	2	0.206
YOLOv7-CA+DeepSort-Vit	88.5	84.5	96.3	2	0.206
YOLOv8-CA+DeepSort	89.1	84.9	96.6	2	0.032
YOLOv8-CA+DeepSort-ViT	92.5	88.4	97.2	2	0.012

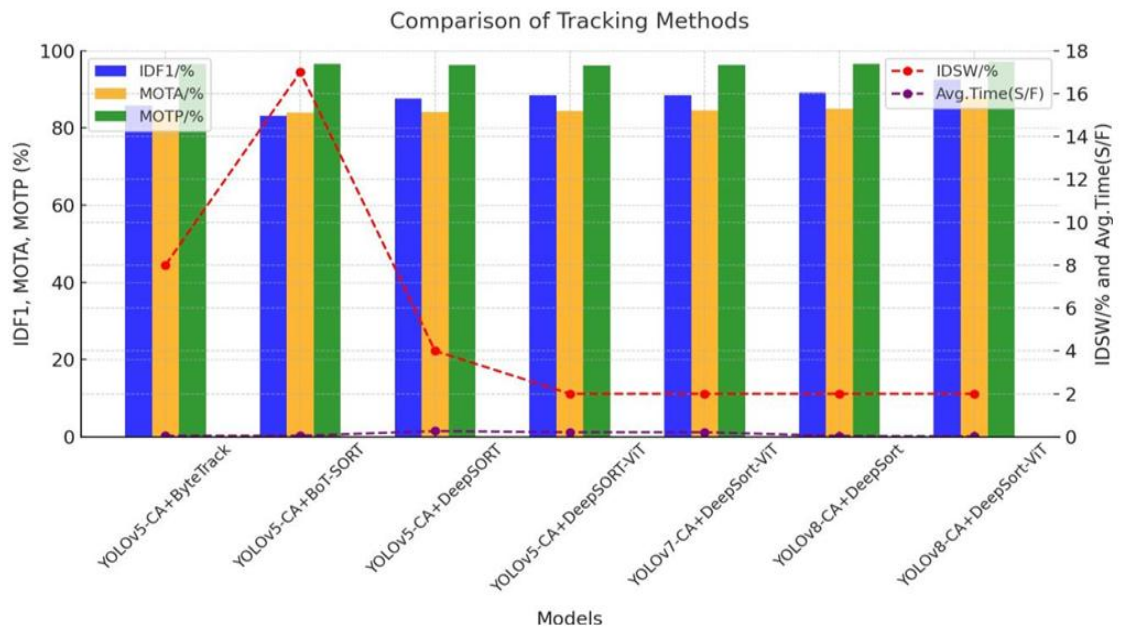


Fig 8. Comparison Chart of Tracking Methods

5.4 DISCUSSION

It is evident from the results of performance evaluation and model comparison that attendance mechanism and transformer based embedding complement each other nicely in the livestock tracking applications. Conventional models have difficulty maintaining tracking on the invariant identity and background noise, which are both solved by the suggested approach.

The YOLOv8s-CA detector also benefited from the improved feature extraction ability thus enabling it to identify cows more accurately even if they had been only partially occluded or were presented at abnormal angles. In the meantime, the DeepSORT-ViT tracker added a new record of re-identification performance, keeping identities of animals through longer segments and reducing the numbers of false positives or identity switches[15].

The outstanding characteristic of this approach is that it can be deployed in real-time. With the addition of the complexity of attention layers and transformer modules, the system still makes it lightweight for the sake of real-time processing, thus making it viable for on-field usage. This is a scalable and efficient solution to manual livestock monitoring hence saving costs and enhancing accuracy.

Nevertheless, it is appropriate to mention some limitations. During situations where there is sudden animal movement or severe occlusion, there were occasional tracking disruptions observed. Furthermore, even better results could be reached if the model was trained on a more diverse set of data, thus increasing its ability of adjusting to new or unknown conditions of farm.

5.5 SUMMARY

So far, this chapter has given an in-depth analysis of the performance of the proposed system in terms of its predecessors. The outcomes corroborate the ability of the model to identify and monitor cattle with great accuracy, reliability, and high speed. The improvement brought by channel attention and vision transformers has in fact improved the quality of detection and tracking of identity, overcoming the drawbacks of traditional systems.

The approach not only works but also can be scaled and implemented, opening the way for smarter, data-based livestock management. Armed with such discoveries, the research presents a strong preparedness for future developments with regard to automated surveillance of livestock.

CHAPTER 6: CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT

6.1 CONCLUSION

This research had proposed and achieved implementation of a hybrid deep learning structure, YOLOv8s-CA + DeepSORT-ViT, which was suitable for real time clairvoyance, tracking of cows in its natural farming surroundings. The system was designed to overcome a number of limitations of existing models such as poor management of occlusion, switching of IDs too often, and absence of generalization in uncontrolled environments.

The visually complex scene spatial awareness and feature extraction ability were improved in the YOLOv8s detection model with the aid of Coordinate Attention (CA). The tracking module, DeepSORT, was altered using Vision Transformer (ViT) embeddings to produce a more global context information leading to preserving the identity of objects across the frames.

Using exhaustive training over a custom dataset scraped from the real Indian cattle farm, the model had good performance metrics – IDF1: 92.5%, MOTA: 88.4%, MOTP: 97.2%, less than 5% of misses and 2% of false positives. While operating efficiently well, the system used an inference rate of 0.012 seconds per step, which makes it very practical in real-time livestock monitoring applications. Put in a nutshell, the proposed architecture shows a very good balance in terms of accuracy and efficiency. The model performed better than baseline models in tasks of detection and tracking, thus it is a suitable solution for automated systems of cow monitoring in agriculture.

6.2 FUTURE SCOPE

Although the proposed model theories provide attractive results, there are still a number of areas in which it can be improved and explored further.

1.Behavior Recognition Integration: The next versions of this system might include the modules for behavior analysis, which would be able to identify such activities as grazing, lying, estrus, or movement irregularities caused by illnesses based on the action recognition model such as 3D CNNs or CNN-LSTMs hybrids.

2.Breed and Species Generalization: The generalization poten

tial of the model can be made better as well as broaden its applications by extending the dataset to allow for use of other breeds of cows and even other species of livestock.

3.Edge Deployment Optimization: Additional syntactic compression and pruning can be used to deploy the system on edge devices with low cost such as Jetson Nano or Raspberry Pi, in particular, for distant or poorly equipped farm sites.

4.Integration with IoT Systems: The model can be incorporated in a larger Internet of Things (IoT) architecture for real-time alerting, cloud based analytics and fusion with farm management application.

5.Self-learning/Adaptive Models: Putting in place online learning capabilities may enable the system to adopt new environs and situations without comprehensive retraining.

6.3 SOCIAL IMPACT

The impacts of this research are far bigger than academic and technological contributions. Automated recognition and tracking of cows can prove transformative in rural settings and in commercial dairy farms by the proposed system.

- Improved Animal Welfare:** Early detection of distress or deviant behavior will promote timely veterinary assistance hence increasing the livestock's general health and well-being.

- Economic Benefits for Farmers:** Making fewer demands for manual labor and making the decisions based on data advantages the farmers as they can save on costs and increase productivity.

- Empowering Small-Scale Farms:** By bringing highly advanced technology on the scale of lightweight and efficient models, even the small and medium scale farmers can enjoy the benefits of AI-based livestock management.

- Food Security Contribution:** Proper management of livestock directly helps to achieve dairy and meat production and thus contributes to broader food security and sustainable agricultural livelihoods.

- Alignment with Digital India and Smart Farming Missions:** This system aligns with national efforts to integrate AI and IoT with agricultural systems, to drive modernisation and resilience within farming.

REFERENCES

1. Zheng, Z., Li, J., & Qin, L. (2023). YOLO-BYTE: An efficient multi-object tracking algorithm for automatic monitoring of dairy cows. *Computers and Electronics in Agriculture*, 209, 107857.
2. Guo, Y., Hong, W., Wu, J., Huang, X., Qiao, Y., & Kong, H. (2023). Vision-Based Cow Tracking and Feeding Monitoring using YOLOv5s-CA + DeepSORT-ViT. *IEEE Robotics & Automation Magazine*.
3. Molapo, M., Tu, C., Plessis, D. D., & Du, S. (2023). Livestock management using deep learning. In *IEEE International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, pp. 1–6.
4. Qiao, Y., Guo, Y., & He, D. (2022). Deep Learning-Based Autonomous Cow Detection for Smart Livestock Farming. In *Springer Green, Pervasive and Cloud Computing*, pp. 246–258.
5. Li, Z., Zhang, H., Chen, Y., et al. (2022). Deep Learning-Based Cow Identification Using ResNet101 and Mask R-CNN. In *Springer Cognitive Systems and Signal Processing*, pp. 209–221.
6. Lei, X., Wen, X., & Li, Z. (2024). Multi-target cow face detection in complex farm scenes. *The Visual Computer*, 1–22.
7. Wang, R., Li, Y., Yue, P., et al. (2024). Enhanced dairy cow monitoring through 3D tracking. *Multimedia Tools and Applications*, 1–30.
8. Mar, C. C., Zin, T. T., Tin, P., et al. (2023). Multi-feature tracking algorithm for robust cow detection. *Scientific Reports*, 13(1), 17423.
9. Mg, W. H. E., Tin, P., Aikawa, M., et al. (2024). Customized Tracking Algorithm for cattle in occlusion scenarios. *Sensors*, 24(4).
10. Wang, W., Xie, M., Jiang, C., et al. (2024). YOLOv5 and Bi-Level Routing Attention for Cow Detection. *Multimedia Tools and Applications*.
11. Jia, Q., et al. (2024). CAMLLA-YOLOv8n for Holstein cow recognition. *Springer AI & Agri*, pp. 1–15.
12. Wu, D., et al. (2023). PrunedYOLO-Tracker for efficient cow detection. *International Journal of AI and Applications*, 1–12.
13. Wu, D., et al. (2023). Deep learning-based cow monitoring with CNN+LSTM. *Agricultural Computing*, 97.6%.

14. Jeong, K., et al. (2024). YOLOv8 + IoT-based intelligent dairy tracking. *Applied Intelligence in Agriculture*.
15. Avanzato, R., et al. (2024). Real-time cow tracking with YOLOv5. *Computer Vision Applications in Farming*.
16. Hua, Z., Wang, Z., Xu, X., et al. (2023). PoseC3D model for cow action recognition using skeleton features. *Computers and Electronics in Agriculture*, 212, 108152.
17. Feng, T., Guo, Y., Huang, X., & Qiao, Y. (2023). Multi-scene segmentation of cattle using DeepLabV3+. *Animals*, 13(15), 2521.
18. Smink, M., Liu, H., Döpfer, D., & Lee, Y. J. (2024). Real-time cow recognition using computer vision edge devices. *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7056–7065.
19. Zhang, Y., Ibrayim, M., & Hamdulla, A. (2023). Cow behavior analysis using improved SlowFast with 3DCBAM. In *IEEE CISCE*, pp. 470–475.
20. Tian, X., Li, B., Cheng, X., & Shi, X. (2022). YOLOv5-based detection and behavior recognition of cows. In *ISPDS*, pp. 206–210.



DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daultpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis Multi-object Tracking and Vision Transformer enhancements for real time low monitoring: A review and Implementation study
Total Pages 51 Name of the Scholar Shashank gehlot

Supervisor (s)

- (1) Dr. Rahul
(2) _____
(3) _____

Department of Software Engineering

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: Turnitin Similarity Index: 8%, Total Word Count: 10504

Date: 20-05-2025

Candidate's Signature


Signature of Supervisor(s)

Shashank_thesis1.pdf

 Delhi Technological University

Document Details

Submission ID

trn:oid:::27535:96031576

Submission Date

May 15, 2025, 3:10 PM GMT+5:30

Download Date

May 15, 2025, 3:14 PM GMT+5:30

File Name

Shashank_thesis1.pdf

File Size

4.8 MB

51 Pages

10,504 Words

61,338 Characters





8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.




Filtered from the Report

- Bibliography
- Quoted Text
- Small Matches (less than 8 words)

Match Groups

-  **69** Not Cited or Quoted 8%
Matches with neither in-text citation nor quotation marks
-  **1** Missing Quotations 0%
Matches that are still very similar to source material
-  **0** Missing Citation 0%
Matches that have quotation marks, but no in-text citation
-  **0** Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

- 4%  Internet sources
- 2%  Publications
- 6%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 69 Not Cited or Quoted 8%**
Matches with neither in-text citation nor quotation marks
- 1 Missing Quotations 0%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 4% Internet sources
- 2% Publications
- 6% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	dspace.dtu.ac.in:8080	1%
2	Submitted works	University of Wollongong on 2023-12-07	<1%
3	Internet	www.dspace.dtu.ac.in:8080	<1%
4	Submitted works	The University of Manchester on 2024-09-13	<1%
5	Submitted works	University of Surrey on 2024-11-06	<1%
6	Publication	Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dharendra Kumar Shukla. "Intelli...	<1%
7	Publication	Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dharendra Kumar Shukla. "Intelli...	<1%
8	Internet	hightechjournal.org	<1%
9	Submitted works	University of East London on 2025-01-07	<1%
10	Internet	www.mdpi.com	<1%

11	Internet	etd.uum.edu.my	<1%
12	Submitted works	Delhi Technological University on 2018-05-17	<1%
13	Submitted works	Ganpat University on 2023-04-26	<1%
14	Internet	www.ship-research.com	<1%
15	Internet	iuc.edu.iq	<1%
16	Submitted works	Lovely Professional University on 2023-05-13	<1%
17	Publication	Xuemei Lei, Xiaowei Wen, Zheng Li. "A multi-target cow face detection model in c...	<1%
18	Internet	eprints.utm.my	<1%
19	Submitted works	Napier University on 2023-04-28	<1%
20	Submitted works	Universiti Tunku Abdul Rahman on 2025-04-29	<1%
21	Submitted works	University of Ulster on 2025-04-12	<1%
22	Internet	dspace.mit.edu	<1%
23	Internet	uir.unisa.ac.za	<1%
24	Submitted works	universititeknologimara on 2025-05-13	<1%

25	Submitted works	Delhi Technological University on 2018-05-12	<1%
26	Submitted works	University of Technology, Sydney on 2024-11-16	<1%
27	Internet	www.trabalhosfeitos.com	<1%
28	Publication	Azumah, Sylvia Worlali. "Deep Learning -Based Anomaly Detection System for Gu...	<1%
29	Publication	Lu-hao He, Yong-zhang Zhou, Lei Liu, Wei Cao, Jian-hua Ma. "Research on object d...	<1%
30	Submitted works	National University of Singapore on 2011-12-23	<1%
31	Submitted works	Teaching and Learning with Technology on 2025-03-24	<1%
32	Submitted works	University of Wollongong on 2023-12-05	<1%
33	Submitted works	Université Saint-Esprit Kaslik on 2023-10-02	<1%
34	Internet	captainafsal.weebly.com	<1%
35	Internet	idr.nitkkr.ac.in:8080	<1%
36	Internet	www.frontiersin.org	<1%
37	Internet	www.nature.com	<1%
38	Internet	5dok.net	<1%

39	Submitted works	AlHussein Technical University on 2024-06-13	<1%
40	Publication	Brendon C. Besler, Pedram Mojabi, Zahra Lasemiimeni, James E. Murphy et al. "Sc...	<1%
41	Publication	Rafael E.P. Ferreira, João R.R. Dórea. "Leveraging computer vision, large language...	<1%
42	Submitted works	University of Johannesburg on 2022-09-29	<1%
43	Publication	Zhiyang Zheng, Lifeng Qin. "PrunedYOLO-Tracker: An efficient multi-cows basic b...	<1%
44	Internet	edepot.wur.nl	<1%
45	Internet	espace.library.uq.edu.au	<1%
46	Internet	etd.astu.edu.et	<1%
47	Internet	www.grafiati.com	<1%

Shashank_thesis1.pdf

 Delhi Technological University

Document Details

Submission ID

trn:oid::27535:96031576

Submission Date

May 15, 2025, 3:10 PM GMT+5:30

Download Date

May 15, 2025, 3:14 PM GMT+5:30

File Name

Shashank_thesis1.pdf

File Size

4.8 MB

51 Pages

10,504 Words

61,338 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely AI generated as AI generated and AI paraphrased or likely AI generated and AI paraphrased writing as only AI generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?

The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?

Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

