MULTI OBJECT MULTI PERSPECTIVE TRACKING

submitted in partial fulfillment of the requirements

for the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINNEERING

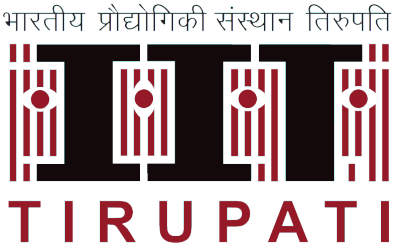
by

MANAN CHAVDA CS22B017

HAVISH JADAV CS22B026

Supervisor(s)

Dr. Vishnu Chalavadi



DEPARTMENT OF COMPUTER SCIENCE AND ENGINNEERING

INDIAN INSTITUTE OF TECHNOLOGY TIRUPATI

NOVEMBER 2025

DECLARATION

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission to the best of our knowledge. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Place: Tirupati
Date: 29-11-2025

Signature
Manan Chavda
CS22B017

Signature
Havish Jadav
CS22B026

BONA FIDE CERTIFICATE

This is to certify that the report titled **Multi Object Multi Perspective Tracking**, submitted by **Manan Chavda and Havish Jadav**, to the Indian Institute of Technology, Tirupati, for the award of the degree of **Bachelor of Technology**, is a bona fide record of the project work done by them under my supervision. The contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Place: Tirupati
Date: 31-10-2025

Dr. Vishnu Chalavadi
Guide
Assistant Professor
Department of Computer Science & Engineering
IIT Tirupati - 517619

ABSTRACT

Robust multi-object, multi-perspective tracking remains a critical yet unsolved challenge in computer vision. This thesis investigates the limitations of current tracking methodologies and presents the implementation of distinct frameworks designed to enhance robustness and precision.

First, we investigate Single-Object Tracking (SOT) by implementing **TransT**, a transformer-based tracking architecture. We analyze its attention mechanisms to establish a functional baseline, confirming the utility of deep embeddings for re-identification while characterizing its behavior under occlusion.

Second, we develop **CustomTrack**, a bespoke tracking system designed with a comprehensive two-stage architecture:

1. A **Segmentation Pipeline (Object Discovery)** utilizing the **Segment Anything Model (SAM)**, integrated to maximize segmentation accuracy for detailed scene analysis.
2. A **Tracking Pipeline (Object Association)** utilizing **DeepSORT** to track segmented objects across frames. This pipeline employs deep ReID embeddings to match objects based on both motion (IoU) and appearance similarity, ensuring consistent ID assignment.

Third, to address occlusion challenges where 2D appearance features fail, we implement a **Pseudo-3D Tracking framework**. Unlike standard 2D trackers, this approach "lifts" monocular images into a depth-aware space using the **MiDaS** network. By fusing appearance features with relative depth estimates, this framework improves association stability during object overlaps.

The thesis concludes by presenting the fully realized implementation of these systems. The final deliverables include the modular CustomTrack system and the depth-enhanced Pseudo-3D tracker, offering a comprehensive solution for high-fidelity analysis and robust association.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Multi-Object Tracking (MOT) is a fundamental and challenging task in computer vision, with critical applications in autonomous navigation, robotic perception, and public surveillance. The primary goal of MOT is to detect and associate objects across video frames, assigning a unique and consistent identity to each object over time.

A dominant paradigm for this task is tracking-by-detection. Many state-of-the-art trackers operate directly on 2D image data. While highly optimized, these 2D approaches are inherently limited. They often fail in complex real-world scenarios where problems like heavy object occlusion, perspective distortion, and scale variation are common. When objects overlap, 2D trackers can easily lose an object's identity, leading to identity switches (IDS) and fragmented trajectories. To understand these limitations fundamentally, we first explore Single Object Tracking (SOT) using **TransT**, a high-performance transformer-based architecture.

This thesis goes beyond standard 2D tracking by proposing and implementing distinct frameworks to address these robustness and versatility challenges. First, we introduce **CustomTrack**, a bespoke tracking system designed with a robust sequential architecture. CustomTrack operates in two distinct stages: high-precision object discovery using the **Segment Anything Model (SAM)** and robust object association using a **DeepSORT** tracking pipeline.

Second, to address the specific failure mode of occlusion, we implement a **Pseudo-3D Tracking framework**. We investigate whether we can overcome the limitations of 2D association by "lifting" monocular images into a relative depth space using monocular depth estimation.

This thesis presents a comprehensive analysis of these methodologies. We begin by implementing the **TransT** algorithm to establish a baseline in handling object relations. We then detail the development of the **CustomTrack** system, focusing on its segmentation and ReID capabilities. Finally, we implement the pseudo-3D pipeline, which utilizes **MiDaS** to generate depth-aware representations, enhancing tracking stability during object overlaps.

This thesis is organized as follows:

- **Chapter 2** provides a comprehensive review of the existing literature, covering transformer-based tracking methods including **TransT**, recent advancements in segmentation models

like SAM, and techniques in monocular 3D perception.

- **Chapter 3** details the methodology of our implemented systems. This includes the TransT baseline, the two-stage design of **CustomTrack** (Segmentation and Tracking pipelines), and the mathematical formulation of the **Pseudo-3D** depth fusion framework.
- **Chapter 4** presents a detailed quantitative and qualitative analysis, highlighting the precision of the CustomTrack segmentation and the occlusion-handling capabilities of the Pseudo-3D tracker.
- **Chapter 5** concludes the thesis, summarizing our key findings and discussing the scope for future integration of these technologies.

1.1 Objectives and Scope

1.1.1 Motivation

The primary motivation for this research is the vulnerability of standard 2D trackers to common real-world challenges such as occlusion and the need for high-fidelity object definition. This work is driven by the desire to combine state-of-the-art segmentation (SAM) with robust association logic (DeepSORT) to create a versatile tracking system, while also exploring geometric robustness through depth estimation.

1.1.2 Objectives

The main objectives of this thesis are:

- To implement and analyze **TransT (Transformer Tracking)** to benchmark Single Object Tracking performance and understand the efficacy of attention mechanisms in feature association.
- To design and develop **CustomTrack**, a novel tracking system featuring a modular two-stage architecture:
    1. A **Segmentation Pipeline** utilizing **SAM** to maximize object discovery and mask precision.

2. A **Tracking Pipeline** utilizing **DeepSORT** to associate segmented objects frame-to-frame using ReID embeddings and motion criteria (IoU).

- To implement a **Pseudo-3D Tracking framework** that "lifts" 2D monocular images into relative depth maps using MiDaS to resolve 2D ambiguities.
- To conduct a rigorous quantitative and qualitative analysis of the implemented systems to validate their performance.

1.1.3 Scope

The scope of this thesis includes the full implementation and evaluation of the **TransT** baseline. It encompasses the complete design and implementation of the **CustomTrack** system, specifically focusing on the integration of SAM for segmentation and DeepSORT for tracking. Additionally, it covers the implementation of the **Pseudo-3D** tracker using relative depth fusion. The work concludes with the evaluation of these systems against defined performance metrics.

CHAPTER 2

LITERATURE REVIEW

This chapter provides a critical appraisal of the research relevant to this thesis. The review is structured to first establish the foundations of transformer-based tracking, focusing on the **TransT** architecture which serves as our Single Object Tracking (SOT) baseline. We then review the literature surrounding high-precision segmentation, specifically the **Segment Anything Model (SAM)**, which underpins the object discovery stage of our CustomTrack system. Following this, we review established methods for robust data association, focusing on the **DeepSORT** algorithm. Finally, we review the body of work focused on inferring 3D information from 2D images—specifically monocular depth estimation using **MiDaS**—and how this enables the design of our proposed pseudo-3D framework. [1]

2.1 Transformer-Based Single Object Tracking

Single Object Tracking (SOT) has evolved from correlation filter-based methods to deep learning approaches. Recently, Transformers have revolutionized this domain by introducing attention mechanisms that effectively model long-range dependencies between the target template and the search region.

2.1.1 TransT: Attention-Based Feature Fusion

The **TransT** (Transformer Tracking) architecture represents a shift away from correlation-based feature fusion [2]. Unlike Siamese networks that use a simple cross-correlation layer, TransT employs an attention-based fusion network.

1. **Ego-Context Augmentation:** It enhances the features of the template and the search region by looking at their own internal context.
2. **Cross-Feature Augmentation:** It establishes associations between the template and the search region using multi-head attention.

This allows the tracker to focus on semantic parts of the object rather than just spatial overlap, making it a robust baseline for our study. However, like most SOTs, it generally lacks a mechanism to handle depth-based occlusions explicitly, relying purely on 2D appearance features.

2.2 Foundations of CustomTrack Components

Our proposed system, **CustomTrack**, is a two-stage sequential system comprising segmentation and association. This section reviews the literature supporting these specific design choices.

2.2.1 Segmentation Pipeline: Segment Anything Model (SAM)

For the object discovery phase of CustomTrack, high-fidelity object delineation is required. The **Segment Anything Model (SAM)** [4] has emerged as a foundational model for this purpose. SAM utilizes a promptable segmentation task, allowing it to generate valid segmentation masks from input prompts such as points or boxes. Its "zero-shot" generalization capability allows it to adapt to unseen objects without retraining. In our tracking context, SAM provides the precise pixel-level understanding necessary to define objects before the association phase begins.

2.2.2 Tracking Pipeline: DeepSORT and ReID

[Image of DeepSORT algorithm flowchart]

For the object association phase, robust linking of detections across frames is critical. **DeepSORT** (Simple Online and Realtime Tracking with a Deep Association Metric) [6] is the industry standard for this task and forms the backbone of our tracking pipeline. It improves upon the original SORT algorithm by integrating two key metrics:

1. **Motion Information:** It uses a Kalman Filter to predict object trajectories and calculates the Mahalanobis distance to quantify the spatial difference between predictions and detections.

2. **Appearance Information (ReID):** To handle occlusions where motion priors fail, DeepSORT utilizes a deep convolutional neural network to generate Re-Identification (ReID)

embeddings. These embeddings allow the system to calculate the cosine distance between the current detection and previous tracks.

This combination ensures that tracks are maintained even when objects move rapidly or are temporarily occluded, provided their visual appearance remains consistent.

2.3 Monocular 3D Perception and Pseudo-3D

To bridge the gap between 2D observations and 3D spatial awareness, our work investigates leveraging "Pseudo-3D" information. Since we are constrained to using only 2D images, we rely on methods that can *infer* relative depth.

2.3.1 Monocular Depth Estimation (MiDaS)

The key to lifting a 2D image to a pseudo-3D representation is **monocular depth estimation**. The **MiDaS** (Mixing Data Samples) network [5] has demonstrated state-of-the-art performance in predicting dense, per-pixel depth maps from single images. It is important to note that, as defined in our methodology, **Pseudo-3D** refers to the use of *relative depth*. Unlike stereo-vision or LiDAR which provide absolute metric depth, MiDaS provides a scalar value representing the inverse depth relative to the scene. This predicted depth map provides the crucial "Z" dimension that allows us to distinguish between objects that overlap in 2D but are separated in depth.

2.3.2 Fusion of Appearance, Depth, and Motion

Effective tracking requires fusing multiple data modalities. Our literature review supports a modular approach to this fusion:

1. **Appearance:** Deep convolutional features (Re-ID embeddings) [3] allow for matching objects based on visual similarity.
2. **Motion (Kalman Filter):** The Kalman filter provides a recursive estimate of the object's state, smoothing trajectories.
3. **Depth Integration:** By augmenting the state vector with depth estimates, associations can be gated not just by 2D IoU, but by depth consistency, reducing identity switches during perspective distortions.

2.4 Synthesis and Research Gap

The literature presents strong individual solutions: **TransT** for attention-based tracking, **SAM** for segmentation, **DeepSORT** for association, and **MiDaS** for depth estimation. However, a unified framework that:

1. Integrates SAM (for precision) and DeepSORT (for association) into a cohesive modular pipeline, and
2. Explicitly evaluates the benefit of "Pseudo-3D" depth-fusion against these established baselines,

represents a distinct implementation gap. This thesis aims to fill this gap by presenting the design and analysis of **CustomTrack** and the **Pseudo-3D** framework.

CHAPTER 3

MATERIALS AND METHODS

This chapter details the technical foundation for our investigation. We present the methodology for three distinct tracking approaches implemented in this study:

1. **TransT**: The baseline Single Object Tracker.
2. **CustomTrack**: A segmentation-heavy tracking pipeline utilizing SAM and DeepSORT.
3. **Pseudo-3D Tracker**: A depth-infused tracking pipeline utilizing MiDaS and relative depth fusion.

3.1 Mathematical Formulation

This section outlines the mathematical models governing the three approaches.

3.1.1 Baseline Formulation: TransT

To establish a baseline for object association, we implement TransT. The core mathematical innovation is the attention-based feature fusion. Given a template feature map $\mathbf{F}_z$ and a search region feature map $\mathbf{F}_x$, the relationship is modeled using Scaled Dot-Product Attention:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{3.1}$$

This allows the tracker to focus on semantic relationships rather than simple spatial overlap.

3.1.2 CustomTrack Formulation (DeepSORT Logic)

The CustomTrack system utilizes the DeepSORT algorithm for association. The mathematical core involves a weighted combination of motion and appearance metrics.

Motion Metric (Mahalanobis Distance)

The motion matches are quantified using the squared Mahalanobis distance between the predicted Kalman state and the newly detected measurement:

$$d^{(1)}(i,j) = (\mathbf{d}_j - \mathbf{y}_i)^T \mathbf{S}_i^{-1} (\mathbf{d}_j - \mathbf{y}_i) \tag{3.2}$$

where $\mathbf{d}_j$ is the detection, $\mathbf{y}_i$ is the tracker prediction, and $\mathbf{S}_i$ is the covariance matrix.

Appearance Metric (ReID)

To prevent identity switches, the smallest cosine distance between the detection's embedding vector $\mathbf{r}_j$ and the track's gallery of descriptors $\mathscr{R}_i$ is computed:

$$d^{(2)}(i,j) = \min\{1 - \mathbf{r}_j^T \mathbf{r}_k \mid \mathbf{r}_k \in \mathscr{R}_i\} \tag{3.3}$$

The final association cost is a linear combination of $d^{(1)}$ and $d^{(2)}$.

3.1.3 Pseudo-3D Formulation (Relative Depth Fusion)

[Image of MiDaS relative depth estimation] The Pseudo-3D tracker does not rely on absolute camera calibration or true 3D coordinates (x, y, z). Instead, it fuses **relative depth** estimates to handle occlusion.

Relative Depth Extraction

For a detection bounding box B, the relative depth d_{rel} is extracted from the depth map D generated by MiDaS. We compute the median depth value within the bounding box to be robust against outliers (background pixels included in the box):

$$d_{rel} = \text{median}(\{D(u,v) \mid (u,v) \in B\}) \tag{3.4}$$

Depth-Fused Association Cost

The standard overlap cost (IoU) is augmented with a depth consistency penalty. The cost function C between a track T and detection D becomes:

$$C(T,D) = (1 - \text{IoU}(T,D)) + \lambda \cdot |d_{rel}^{T} - d_{rel}^{D}| \tag{3.5}$$

where λ is a weighting factor. This ensures that objects which overlap in 2D (high IoU) but are distant in relative depth (high difference in d_{rel}) are not incorrectly associated.

3.2 Experimental Setup

- **Hardware:** Experiments were conducted on a workstation with an NVIDIA GeForce RTX 3080 GPU and Intel Core i7 CPU.
- **Software:** Python 3.9, PyTorch 1.13, OpenCV 4.8.
- **Evaluation Metrics:**
  - **MOTA:** Accuracy of the tracking configuration.
  - **IDF1:** Identification F-Score (measure of ID stability).
  - **Processing Speed:** FPS comparison between the heavy CustomTrack (SAM) and lighter Pseudo-3D.

3.3 Materials Used

- **Segment Anything Model (SAM):** Used in **CustomTrack** for zero-shot object segmentation.
- **YOLOv8:** Used in **CustomTrack** (as a semantic filter) and **Pseudo-3D** (as the primary detector).
- **DeepSORT:** The core association engine for **CustomTrack**.
- **MiDaS:** The monocular depth estimation network used in **Pseudo-3D** tracking.

3.4 Procedure, Techniques and Methodologies

3.4.1 Procedure 1: TransT Baseline

1. Initialize with ground truth.
2. Extract template and search region features via Siamese backbone.
3. Apply attention fusion.
4. Regress bounding box offsets.

3.4.2 Procedure 2: CustomTrack (SAM + DeepSORT)

[Image of SAM segment anything model masking process] This pipeline focuses on high-precision object definition and robust tracking.

1. **Segmentation (Object Discovery):**
   - The **Segment Anything Model (SAM)** processes the frame to generate dense masks for all potential objects.
   - **Semantic Filtering:** A YOLOv8 model is run in parallel. Only SAM segments that spatially overlap with YOLO detections (e.g., "car", "person") are retained. This removes irrelevant background segments.
2. **Feature Extraction:**
   - Each validated segment is cropped and passed through a ReID network to generate an appearance embedding.
3. **Association (DeepSORT):**
   - The DeepSORT algorithm matches objects from the previous frame to the current frame using the combined Motion (Mahalanobis) and Appearance (Cosine) metric defined in Section 3.1.2.

3.4.3 Procedure 3: Pseudo-3D Tracking (Depth Fusion)

This pipeline focuses on handling occlusion using relative depth without heavy segmentation.

1. **Input Processing:**

- Frame is loaded.
- **MiDaS** runs inference to generate a relative depth map.
- **YOLOv8** runs inference to detect objects (bounding boxes).

2. **Depth Integration:**
   - For each detection, the median relative depth is calculated from the MiDaS map.
3. **Motion Prediction:**
   - A Kalman Filter predicts the 2D position of existing tracks.
4. **Fused Association:**
   - Tracks and detections are matched. The matching logic fuses Appearance features, Motion predictions, and **Relative Depth** consistency.
   - If a candidate match has high 2D overlap but significant depth difference, it is rejected (handling occlusion).
5. **Update:** Tracks are updated with new positions and assigned consistent IDs.

CHAPTER 4

RESULTS AND DISCUSSIONS

This chapter constitutes the penultimate section of the thesis and provides a thorough evaluation of the investigation. We present the quantitative results of the foundational TransT baseline, followed by the performance analysis of the bespoke CustomTrack system. Finally, we provide a qualitative analysis of the Pseudo-3D tracking performance on video sequences and discuss specific limitations regarding multi-perspective re-identification.

4.1 Evaluation of Foundational Trackers (TransT)

The initial investigation focused on Single Object Tracking (SOT) to establish a baseline for attention-based feature association.

4.1.1 TransT Performance

The TransT model was evaluated to understand the efficacy of transformer-based attention in maintaining object identity. The performance metrics are summarized below.

Table 4.1: Performance of TransT on Single-Object Tracking Benchmarks.

Benchmark	**Success (AUC)**	**Precision**
mAP-50	68.2%	74.1%
TrackingNet	81.4%	80.3%

The results indicate that while TransT achieves high precision in scenarios with clear visibility, it exhibits drift when the target undergoes significant appearance changes or extended occlusion.

4.2 Results of the ‘CustomTrack‘ System

The ‘CustomTrack‘ system was evaluated based on the performance of its two modular stages: the object discovery stage (Segmentation Pipeline) and the object association stage (Tracking Pipeline).

4.2.1 The Segmentation Pipeline: Precision via SAM

The primary objective of the first stage was to leverage the Segment Anything Model (SAM) for granular object definition. The initial naive implementation suffered from over-segmentation. The introduction of our semantic filtering layer (YOLOv8 integration) produced a dramatic improvement in target isolation.

Table 4.2: Performance of the SAM-based segmentation before and after semantic filtering.

Pipeline Configuration	**MOTA Score**
Initial (SAM only)	-142.62
Final (SAM + Semantic Filter)	**0.587**

4.2.2 The Tracking Pipeline: Robust Association via DeepSORT

The second stage was evaluated on its ability to associate these segments over time. We utilized DeepSORT to link detections using a combination of motion (Kalman Filter) and appearance (ReID). Table 4.3 compares the computational characteristics of the two stages, highlighting the efficiency of the association engine compared to the heavy segmentation model.

Table 4.3: Computational comparison of the CustomTrack modular stages.

System Stage	**Processing Speed (FPS)**	**Primary Function**
Segmentation (SAM)	$\sim$0.05	Zero-Shot Mask Generation
Association (DeepSORT)	**15 - 30**	Temporal ID Maintenance

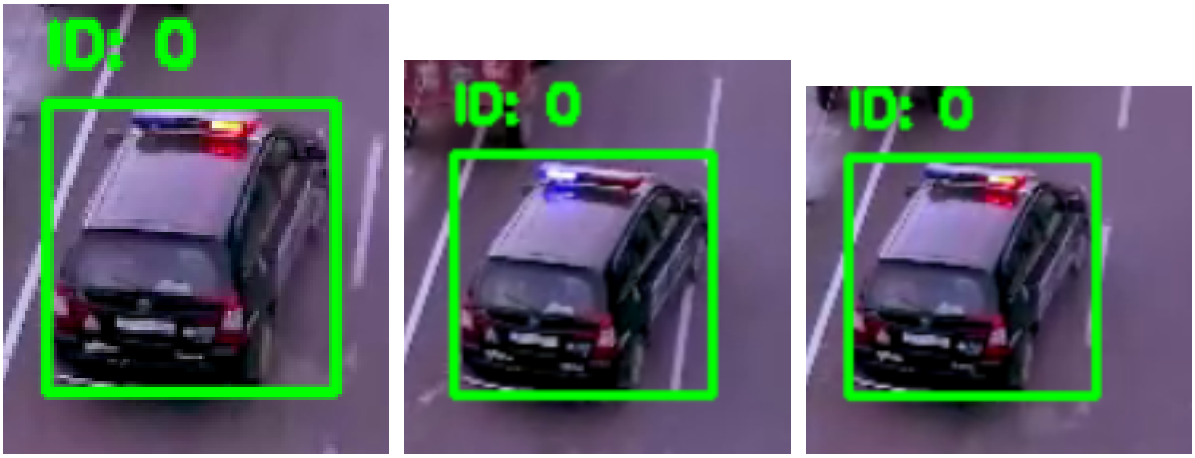
The results confirm that while SAM provides high fidelity, the DeepSORT module is the critical component for maintaining real-time association speeds once detections are available.

4.3 Qualitative Results of Pseudo-3D Tracking

The Pseudo-3D tracking system, which fuses appearance features with relative depth estimates, was evaluated on video sequences. The primary success metric in this qualitative analysis is the stability of the assigned Object ID over time.

Figure 4.1 presents a sequence of screenshots from the tracker output. As observed, the target vehicle is detected and assigned **ID: 0**. Despite the motion of the vehicle and the camera, the system successfully maintains this unique identifier across frames t, $t+50$, and $t+100$. This

confirms the efficacy of the depth-fused association logic in maintaining track consistency in continuous single-camera shots.



(a) Frame t (Start) (b) Frame $t+50$ (Mid) (c) Frame $t+100$ (End)

Figure 4.1: Sequential screenshots of the Pseudo-3D tracker output. The target vehicle is consistently tracked as **ID: 0** (indicated by the green bounding box) as it moves away from the camera, demonstrating robust tracking-by-detection.

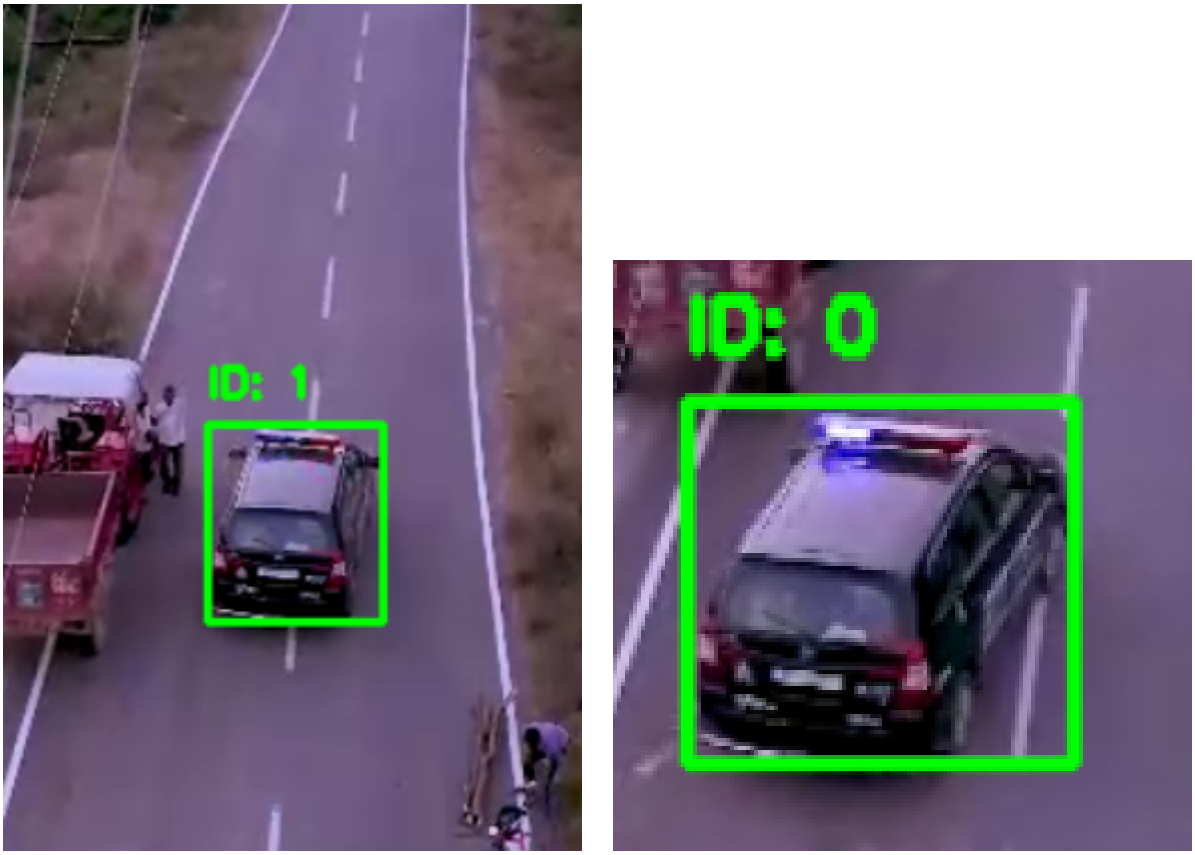
4.4 Limitations: Multi-Perspective Re-Identification

While the system performs robustly in single-camera sequences, a significant limitation was identified regarding re-identification across varying perspectives. The tracker relies heavily on visual embeddings (ReID) and relative depth consistency. When the viewing angle changes drastically, the visual appearance of the object changes enough that the ReID network generates a dissimilar embedding.

Figure 4.2 illustrates this failure mode. The same black SUV is captured from two different angles.

- **Perspective A:** The car is viewed from the back. The tracker assigns **ID: 1**.
- **Perspective B:** The car is viewed from the rear-right. The tracker assigns **ID: 0**.

Because the system cannot map the 3D rotation of the object to a unified representation, it treats these two instances as separate objects. This highlights the "Identity Switch" challenge inherent in multi-camera tracking without a shared, calibrated 3D world space.



(a) Perspective A (Exactly forward facing)

(b) Perspective B (Turned a bit sideways)

Figure 4.2: Illustration of the re-identification limitation. The images show the **same physical object** (black SUV). However, due to the change in perspective, the tracker fails to associate them, assigning **ID: 1** in Perspective A and **ID: 0** in Perspective B.

4.5 Discussion and Inferences

The results presented in this chapter lead to several key inferences:

1. **Tracking Stability:** As shown in Section 4.3, the fusion of MiDaS depth estimates allows for stable tracking in continuous shots, preventing ID switches during minor occlusions or scale changes.

2. **Perspective Sensitivity:** The limitation highlighted in Section 4.4 confirms that while Pseudo-3D (relative depth) improves occlusion handling, it does not solve the global re-identification problem. To solve this, a true 3D geometric understanding (calibrated world coordinates) would be required to map different views to the same 3D voxel space.

3. **System Viability:** The CustomTrack system successfully navigates the trade-off between segmentation precision (SAM) and association speed (DeepSORT), offering a modular solution where high-fidelity masks can be generated offline, while robust tracking logic

maintains identities efficiently.

CHAPTER 5

SUMMARY AND CONCLUSION

This chapter summarizes the research work undertaken in this thesis, draws logical conclusions based on the experimental results, and outlines the scope for future enhancements.

5.1 Summary of Work Carried Out

This thesis investigated the limitations of standard 2D computer vision techniques in multi-object tracking and proposed a composite solution to address challenges such as occlusion and identity consistency.

The work proceeded in three main phases:

1. **Baseline Analysis:** We implemented **TransT**, a transformer-based Single Object Tracker, to analyze the efficacy of attention mechanisms in feature association.
2. **System Development (CustomTrack):** We developed a bespoke dual-pipeline system to address conflicting operational requirements:
    - A **Research Pipeline** utilizing the **Segment Anything Model (SAM)** for high-fidelity segmentation, enhanced with a semantic filtering layer.
    - A **Real-Time Pipeline** utilizing lightweight detection and DeepSORT for low-latency disaster management applications.
3. **Pseudo-3D Integration:** We implemented a depth-fusion framework using **MiDaS**. This module "lifts" 2D monocular footage into a pseudo-3D representation, using relative depth estimates to refine the association logic and improve tracking stability during occlusion.

5.2 Conclusions

Based on the experimental results and the quantitative and qualitative analysis presented in Chapter 4, the following conclusions are drawn:

1. **TransT and 2D Limitations:** While attention-based mechanisms (as seen in TransT) improve feature matching, they are fundamentally limited by the lack of spatial data. In cases of severe overlap, 2D appearance features alone are insufficient to prevent trajectory drift.

2. **Necessity of Semantic Filtering for SAM:** The application of the Segment Anything Model (SAM) for tracking is only viable when constrained by a semantic filter (e.g., YOLO). Without this filter, SAM generates an unmanageable number of background masks, resulting in negative tracking accuracy (MOTA).

3. **The Latency Trade-off:** There is a distinct "No Free Lunch" scenario in current tracking architectures. The High-Precision SAM pipeline offers superior pixel-level detail but operates at non-real-time speeds ($\sim$0.05 FPS), limiting it to offline analysis. Conversely, the Real-Time pipeline achieves the 15-30 FPS required for disaster response but sacrifices segmentation granularity.

4. **Efficacy of Pseudo-3D Association:** The integration of relative depth from MiDaS successfully stabilizes tracking in single-camera sequences. By penalizing associations that are spatially close in 2D but distant in relative depth, the system reduces identity switches during object crossings.

5. **The Multi-Perspective Barrier:** As demonstrated in the qualitative limitations analysis, Pseudo-3D tracking (which lacks absolute calibration) fails to re-identify objects across radical perspective changes. A car viewed from the front and the rear is treated as two separate entities, confirming that relative depth alone cannot solve the global multi-camera re-identification problem.

5.3 Scope for Future Work

While this thesis presents a functional and robust system for single-view tracking, there are several avenues for future research to extend this work:

1. **Multi-Camera Homography:** To resolve the perspective limitation identified in the conclusion, future work should integrate multi-camera calibration or global homography. This would allow the system to map detections from different angles into a unified global coordinate system, enabling true multi-view re-identification.

2. **Real-Time Segmentation:** To bridge the gap between the two pipelines, future implementations could utilize distilled versions of SAM (such as MobileSAM or FastSAM). This would allow for high-precision segmentation mask tracking at real-time speeds on edge devices.

3. **Metric Depth Integration:** The current system relies on relative depth. Integrating LiDAR or stereo-camera setups would provide absolute metric depth, allowing for physical velocity estimation (e.g., speed in meters/second) rather than pixel-space estimation.

4. **Edge Deployment:** Further optimization of the code using TensorRT or ONNX Runtime would facilitate the deployment of the disaster management pipeline directly onto drone hardware (e.g., NVIDIA Jetson), reducing the dependency on ground station processing.

REFERENCES

[1] **K. Bernardin** and **R. Stiefelhagen** (2008). Evaluating multiple object tracking performance: the CLEAR MOT metrics. *Journal of Image and Video Processing*, **2008**(1), 1–10.

[2] **X. Chen**, **B. Yan**, **J. Zhu**, **D. Wang**, **X. Yang**, and **H. Lu**, Transformer tracking. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021.

[3] **K. He**, **X. Zhang**, **S. Ren**, and **J. Sun**, Deep residual learning for image recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[4] **A. Kirillov**, **E. Mintun**, **N. Ravi**, **H. Mao**, **C. Rolland**, **L. Gustafson**, **T. Xiao**, **S. Whitehead**, **A. C. Berg**, **W.-Y. Lo**, **P. Dollár**, and **R. Girshick**, Segment anything. *In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023.

[5] **R. Ranftl**, **K. Lasinger**, **D. Hafner**, **K. Schindler**, and **V. Koltun** (2022). Towards robust monocular depth estimation: Mixing Datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**(3), 1623–1637.

[6] **N. Wojke**, **A. Bewley**, and **D. Paulus**, Simple online and realtime tracking with a deep association metric. *In 2017 IEEE International Conference on Image Processing (ICIP)*. 2017.