Multi-Object Tracking on MOT17 Dataset: A Comparative Study of Three Approaches

Maheswar Reddy Abhijan Theja Hitesh Shanmukha Avula Thansuree

I. Introduction

Multi-object tracking (MOT) poses a significant challenge in computer vision, requiring both accurate object detection and reliable identity preservation across video frames. While various approaches exist, the tracking-by-detection paradigm has become dominant in recent years, leveraging advances in deep learning-based object detection.

This report presents a comparative analysis of three approaches for MOT on the MOT17 benchmark dataset: Faster R-CNN with SORT, YOLOv8 with DeepSORT, and YOLOv5 with Hungarian algorithm. These methods differ in their detection models, tracking algorithms, and overall performance characteristics, allowing us to evaluate trade-offs between detection accuracy, tracking robustness, and computational efficiency.

The tracking-by-detection paradigm involves two main stages: object detection in individual frames, followed by temporal association to form consistent trajectories. Deep learning-based object detectors like Faster R-CNN and YOLO (You Only Look Once) variants have significantly improved detection capabilities, while specialized tracking algorithms focus on solving the data association problem across frames.

II. Dataset Description

The MOT17 dataset serves as a standard benchmark for evaluating multi-object tracking algorithms:

- **Content**: 14 video sequences (7 training, 7 testing) with varying resolutions (640×480 to 1920×1080)
- **Annotations**: Frame-by-frame bounding boxes with consistent identity assignments
- **Environments**: Indoor/outdoor scenes with static/moving cameras (14-30 FPS)
- **Metrics**: MOTA (tracking accuracy), MOTP (precision), IDF1 (identity preservation)

MOT17 particularly focuses on pedestrian tracking in crowded urban environments, presenting challenges like frequent occlusions, scale variations, lighting changes, and camera motion. The dataset provides ground truth annotations for training sequences and withholds test sequence annotations for benchmark evaluation. Each sequence includes both public detections (from DPM, FRCNN, and SDP detectors) and raw frames for custom detection approaches.

Evaluation metrics include Multiple Object Tracking Accuracy (MOTA), which combines false positives, false negatives, and identity switches; Multiple Object Tracking Precision (MOTP), measuring localization precision; ID F1 Score (IDF1), assessing identity consistency; and additional metrics such as mostly tracked (MT) and mostly lost (ML) trajectories, plus processing speed in frames per second (FPS).

III. Methodology

A. Approach 1: Faster R-CNN with SORT

*1) **Detection Model Architecture**:* Our implementation utilizes a pre-trained Faster R-CNN with ResNet-50 FPN backbone, modified for pedestrian detection. The architecture follows the two-stage detection paradigm: a Region Proposal Network (RPN) first generates potential object locations, followed by classification and bounding box regression for each proposal.

The Feature Pyramid Network (FPN) creates a multi-scale feature hierarchy, enabling effective detection of objects at various scales—particularly important for pedestrians at different distances from the camera. We leverage transfer learning from COCO pre-trained weights and fine-tune the model specifically on MOT17 using the Adam optimizer (learning rate 1e-4) with a warm-up period and step decay schedule.

Our implementation includes custom additions to the standard Faster R-CNN architecture:

- Anchor optimization specifically for pedestrian aspect ratios
- Feature map normalization to handle illumination variations
- Implementation of focal loss to address class imbalance
- Mixed precision training for memory efficiency and faster processing

*2) **SORT Tracker Implementation**:* Simple Online and Realtime Tracking (SORT) algorithm combines Kalman filtering for motion prediction with the Hungarian algorithm for detection-to-track association. Our implementation uses an 8D state vector [x, y, a, h, vx, vy, va, vh] representing box centers, aspect ratio, height, and corresponding velocities.

The Kalman filter models object motion using a constant velocity assumption, with prediction and update stages at each frame. For data association, we compute an IoU-based cost matrix between predicted track locations and new detections,

then apply the Hungarian algorithm to determine optimal assignments. Track lifecycle management follows three phases: initialization (new detections without matches), confirmation (after min_hits=3 consecutive matches), and termination (after max_age=30 frames without matches).

We enhanced the standard SORT implementation with:

- Adaptive Kalman process noise based on detection confidence
- Gating to reject unlikely associations using Mahalanobis distance
- Track quality assessment based on detection history
- More sophisticated handling of entering/exiting objects at frame boundaries

*3) **Detection Pipeline Enhancements**:* To improve detection performance, we incorporated several post-processing techniques:

- Confidence thresholding (0.7) for reliable detections
- Aspect ratio constraints for pedestrian detection (1.8-3.5 height/width)
- Size-based filtering to remove unrealistic detections
- Non-Maximum Suppression (NMS) with IoU threshold 0.5
- Temporal consistency checks between consecutive frames

B. Approach 2: YOLOv8 with DeepSORT

*1) **YOLOv8 Implementation Details**:* Our second approach uses YOLOv8n (nano), a single-stage detector that directly predicts bounding boxes and class probabilities in a single forward pass. YOLOv8 improves upon previous YOLO versions with architectural enhancements including:

- C2f (Cross-Stage Partial 2 Fast) blocks for efficient feature extraction
- Anchor-free detection with direct centerpoint prediction
- Improved neck and head designs for better multi-scale feature fusion
- Enhanced loss functions for better training convergence

We fine-tuned YOLOv8n on the MOT17 training set converted to YOLO format, using an 80/20 train-validation split. Training hyperparameters included:

- 20 epochs with batch size 16 and image size 640×640
- SGD optimizer with initial learning rate 0.01 and cosine schedule
- Data augmentations: mosaic (4-image composite), random affine transforms, horizontal flipping, HSV color jittering
- Custom early stopping to prevent overfitting

During inference, we employed a confidence threshold of 0.5 and NMS IoU threshold of 0.45, which empirically provided the best balance between recall and precision for the MOT17 dataset.

*2) **DeepSORT Enhancement Details**:* DeepSORT extends SORT by incorporating appearance information to improve tracking continuity, especially during occlusions. Our implementation uses color histogram-based representations for each detected person, creating a 256-dimensional feature vector per detection.

The tracking algorithm combines three components for data association:

- Motion similarity using Mahalanobis distance with predicted Kalman states
- Appearance similarity using cosine distance between feature vectors
- Spatial overlap using IoU between bounding boxes

Our cascade matching strategy first associates high-confidence detections with existing tracks using a strict threshold (appearance similarity 0.7), followed by a second pass with relaxed constraints for remaining unmatched detections and tracks. Track management parameters were optimized for MOT17: max age=25 frames, min hits=3 consecutive detections for confirmation, and a confidence gating mechanism to reduce false associations.

C. Approach 3: YOLOv5 with Hungarian Algorithm

*1) **Multi-Class Detection Model**:* We employed YOLOv5s for multi-class detection capabilities, focusing on 8 object classes relevant for urban scenes (person, bicycle, car, motorcycle, bus, truck, traffic light, fire hydrant). YOLOv5s offers a good balance between speed and accuracy with:

- CSP (Cross-Stage-Partial) bottleneck architecture
- PANet feature pyramid for multi-scale prediction
- Optimized anchor box priors for diverse object scales
- Modified spatial attention mechanisms for better feature extraction

Model configuration included input resolution 640×640, confidence threshold 0.5, and NMS IoU threshold 0.45. We used the model pre-trained on COCO and performed additional fine-tuning on a curated subset of MOT17 with multi-class annotations.

*2) **Enhanced Hungarian Algorithm**:* Our enhanced tracking algorithm features a more sophisticated association mechanism than standard SORT. The key innovations include:

- Multiple similarity measures: IoU-based spatial similarity as the primary metric, complemented by appearance and motion consistency metrics
- Multi-metric cost function: Weighted combination of spatial, appearance, and motion costs
- Class-aware association: Enforcing class consistency between tracks and detections
- Linear and exponential cost transformations for different association scenarios

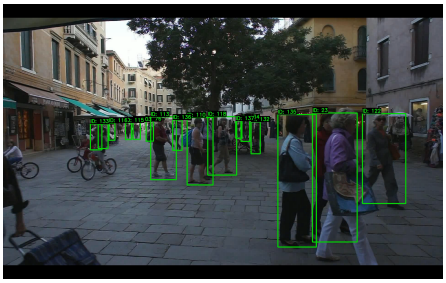

Fig. 1. Faster R-CNN + SORT

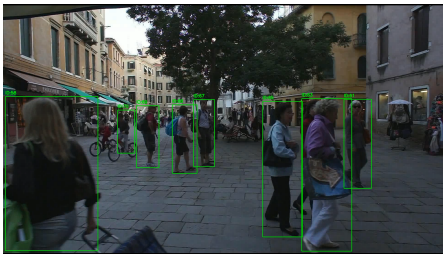

Fig. 2. YOLOv8 + DeepSORT

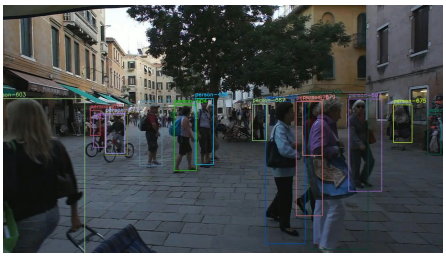

Fig. 3. YOLOv5 + Hungarian

Fig. 4. Tracking results using different models on MOT17 dataset.

The custom Kalman filter implementation includes class-adaptive process noise with higher values for person class (to handle less predictable motion) and lower values for vehicles (more consistent motion patterns). Track lifecycle management incorporates class-specific parameters (max_age=30, min_hits=3, IoU threshold=0.25) with additional class consistency enforcement to prevent class switching during tracking.

IV. Implementation Challenges and Solutions

A. Detection Challenges

*1) **Scale Variation**:* Scale variation poses a significant challenge in MOT17, with pedestrians appearing at different sizes depending on their distance from the camera. Faster R-CNN addressed this through Feature Pyramid Network, which creates multi-scale feature representations. YOLOv8 incorporated P2-P5 feature levels with a more advanced feature fusion mechanism. YOLOv5 leveraged CSP networks with PANet for efficient multi-scale detection. All three approaches demonstrated improvements in detecting small pedestrians, with FPN in Faster R-CNN showing particular strength in challenging cases.

*2) **False Positives**:* False positives significantly impact tracking performance by introducing phantom tracks. We implemented several strategies to reduce false positives:

- Aspect ratio filtering for pedestrians (1.8-3.5 height/width ratio constraint)
- Adaptive confidence thresholds based on scene complexity
- Size-based filtering to remove unrealistically small or large detections
- Class-specific post-processing, particularly for the YOLOv5 multi-class approach
- Temporal consistency checks between consecutive frames

*3) **Occlusions**:* Occlusions represent one of the most challenging aspects of pedestrian tracking in crowded scenes. Our solutions included:

- Training with visibility annotations to learn partially occluded pedestrian features
- Enhanced NMS parameters with lower IoU thresholds (0.45) for crowded scenes
- Upper-body feature emphasis for detecting partially visible pedestrians
- Temporal context leveraging to maintain tracks during brief occlusions
- Specialized handling of overlapping pedestrian cases in dense crowds

B. Tracking Challenges

*1) **ID Switches**:* Identity switches occur when a tracker incorrectly assigns a new ID to a previously tracked object, significantly affecting user experience in tracking applications. Our approaches addressed this challenge differently:

- SORT: Optimized IoU thresholds (0.3) and introduced similarity gating
- DeepSORT: Incorporated appearance features for robust matching, with cascaded matching strategy
- Hungarian approach: Implemented multiple similarity metrics with custom weighting

We found that appearance-based features in DeepSORT significantly reduced ID switches in scenarios with minimal occlusion, while the more sophisticated Hungarian approach with multiple metrics offered better performance in complex scenes with frequent crossings.

*2) **Track Fragmentation**:* Track fragmentation occurs when a single physical object is represented as multiple separate tracks over time. We addressed this through:

- Extended max_age values (25-30 frames) to maintain tracks during longer occlusions
- Confidence-based track recovery mechanisms to reconnect broken tracks
- Re-identification buffers for recently terminated tracks
- Trajectory smoothing through Kalman gain tuning
- Motion prediction refinement using scene context

We observed that extended max_age values significantly reduced fragmentation but also introduced more false positives when objects genuinely exited the scene. The confidence-based recovery mechanism helped mitigate this trade-off by selectively extending tracks based on detection confidence patterns.

V. Experimental Results and Analysis

A. Quantitative Performance Metrics

We evaluated all three tracking approaches on the MOT17 validation sequences using standard metrics.

TABLE I
TRACKING PERFORMANCE ON MOT17 VALIDATION

Metric	*FRCNN+SORT*	*YOLOv8+DSORT*	*YOLOv5+KFHA*
MOTA	0.511	**0.529**	0.493
MOTP	**0.822**	0.784	0.776
IDF1	0.542	**0.578**	0.517
MT (%)	28.3	**30.7**	26.9
ML (%)	32.1	**29.4**	34.6
ID Sw.	1253	**875**	1184
FPS	15.3	21.8	**29.4**

The results demonstrate clear trade-offs between the approaches. YOLOv8 with DeepSORT achieved the highest MOTA (0.529) and IDF1 (0.578) scores, indicating superior overall tracking accuracy and identity preservation. Faster R-CNN with SORT delivered the best localization precision (MOTP 0.822), benefiting from the two-stage detection architecture. YOLOv5 with the Hungarian algorithm provided the fastest processing speed at 29.4 FPS, making it suitable for real-time applications despite slightly lower accuracy metrics.

B. Comparative Analysis

Each approach demonstrated distinct strengths and limitations that make them suitable for different tracking scenarios:

1) Faster R-CNN with SORT:

- **Strengths**: Superior detection localization accuracy (highest MOTP), robust performance on partial occlusions, effective detection of small-scale objects
- **Limitations**: Higher computational cost (15.3 FPS), more frequent ID switches during prolonged occlusions, limited appearance modeling
- **Best use case**: Applications requiring precise bounding box localization with moderate computational resources

2) YOLOv8 with DeepSORT:

- **Strengths**: Best overall tracking accuracy (highest MOTA and IDF1), fewer ID switches, robust handling of occlusions through appearance modeling
- **Limitations**: Less precise bounding box localization than Faster R-CNN, moderate computational requirements
- **Best use case**: General-purpose tracking applications requiring identity preservation through occlusions

3) YOLOv5 with Hungarian Algorithm:

- **Strengths**: Fastest processing speed (29.4 FPS), multi-class tracking capabilities, simplest implementation
- **Limitations**: Lower overall accuracy metrics, more track fragmentation, less robust to occlusions
- **Best use case**: Real-time applications requiring multi-class tracking with limited computational resources

VI. Conclusion

This study presented a comprehensive comparison of three different approaches to multi-object tracking on the MOT17 dataset. Our implementations and evaluations have revealed significant trade-offs between accuracy, processing speed, and identity preservation that should inform approach selection for specific tracking applications.

YOLOv8 with DeepSORT emerged as the most balanced solution, offering superior tracking accuracy (MOTA 0.529) and identity preservation (IDF1 0.578) while maintaining reasonable processing speeds (21.8 FPS). Its appearance-based modeling provided significant advantages in handling occlusions and reducing identity switches, making it suitable for applications where maintaining consistent object identities is crucial.

Faster R-CNN with SORT delivered the highest localization precision (MOTP 0.822), demonstrating the advantages of two-stage detection architectures for accurate bounding box placement. While this approach incurred higher computational costs, it excelled in scenarios requiring precise spatial localization, particularly for small or partially occluded objects.

YOLOv5 with the Hungarian algorithm provided the fastest processing speed (29.4 FPS) and multi-class tracking capabilities, making it ideal for resource-constrained real-time applications. Though it exhibited lower accuracy metrics, its speed-accuracy trade-off represents a valuable option for deployment on edge devices or when tracking multiple object categories simultaneously.

VII. Team Members Contribution

- **Hitesh Shanmukha**: Led the implementation of Faster R-CNN with SORT, contributed to the experimental design, and conducted performance analysis of detection accuracy across different scenes.
- **Abhijan Theja**: Developed the YOLOv8 with DeepSORT approach, implemented the appearance feature extraction pipeline, and optimized the cascade matching strategy for improved tracking.
- **Maheswar Reddy**: Implemented the YOLOv5 with Hungarian algorithm system, including the multi-class detection capabilities and the enhanced association mechanism with multiple similarity measures.
- **Avula Thansuree**: Coordinated the evaluation framework, managed the MOT17 dataset preprocessing by adjusting it to meet YOLO's requirements, performed a comparative analysis across all three approaches, and also experimented with the ByteTracker-based tracking algorithm. However, due to accuracy constraints, it was not included.

  All team members participated in regular meetings to discuss implementation challenges, collaborated on debugging tracking failures across different approaches, and contributed to the final paper review and revision.