Hybrid Transformer–YOLO Framework for Real-Time Object Detection and Tracking in Complex Surveillance Environments

Dr.Jayaram C V[1], Dr.K. Sudhakar[2]

[1]PostDoctoral Researcher, *Lincoln University College Malaysia;* [2]*Artificial Intelligence & Data Science, Nitte Meenakshi Institute of Technology, Nitte (Deemed to be University), Bengaluru, Karnataka, India.*

Pdf.jayaramscv@lincoln.edu.my; ksudhakar.cs@gmail.com

Abstract: Real-time object detection and tracking remain difficult in complex surveillance settings, where dense crowds, occlusion, variable illumination, and fast-moving targets are common. YOLO-based detectors run quickly but struggle to model long-range spatial dependencies, whereas transformer architectures capture global context at considerable computational expense. This paper introduces the Hybrid Transformer–YOLO (HT-YOLO) framework, which pairs a lightweight YOLO backbone with a compact transformer encoder through cross-attention fusion, and couples this detector to a tracking stage built on Kalman-filter motion prediction and appearance-based re-identification (Re-ID). Evaluated on the MOT17 and MOT20 benchmarks, HT-YOLO runs at 42 FPS on an RTX 4090 GPU while delivering gains of 4.6 points in mAP@0.5 and 6.1 points in MOTA over a YOLOv8 baseline. The results highlight the contribution of the cross-attention fusion module and the Re-ID-aware tracker, especially under occlusion, and suggest that hybrid CNN–transformer designs can support accurate, low-latency detection and tracking for surveillance applications.

Keywords: Object Detection; Multi-Object Tracking; YOLO; Vision Transformer; Cross-Attention Fusion; Re-Identification; Surveillance; Real-Time Systems.

Introduction

Video surveillance underpins public safety, traffic monitoring, retail analytics, and infrastructure protection, and each of these applications depends on multi-object detection and tracking that stays reliable under occlusion, crowding, scale variation, and changing illumination. YOLO-family convolutional detectors are the dominant choice for real-time use, yet their limited receptive field makes it hard to localize distant or partially hidden objects, particularly in dense scenes. Vision transformers and DETR-style models close this gap through self-attention, which extracts global context and improves robustness to occlusion, but the quadratic cost of full self-attention makes such models difficult to run in real time on surveillance hardware.

To close this gap, we propose HT-YOLO, a framework that augments a YOLO feature extractor with a lightweight transformer encoder applied to down-sampled tokens and merges the two representations through cross-attention prior to detection. A tracking stage is coupled directly to the detector, combining Kalman filtering with Re-ID embeddings to preserve object identity through brief periods of occlusion.

Related Work

YOLO-family detectors cast detection as a single-stage regression problem, offering a strong balance of speed and accuracy but remaining constrained by the local receptive field of convolution in dense, occluded scenes. DETR-style transformers instead rely on self-attention and learned object queries to capture global context, which improves robustness to small objects and occlusion at the cost of substantially higher per-frame latency. Hybrid CNN–transformer designs retain a convolutional backbone for efficiency and add transformer blocks at coarser resolutions; such designs have proven competitive on classification and segmentation but remain comparatively unexplored for real-time surveillance detection and tracking. Tracking-by-detection methods typically pair appearance embeddings for re-

identification with motion models such as Kalman filters. In this work, tracking is tied directly to the hybrid detector's aggregated features rather than treated as an independent stage.

Key Contribution

This work presents HT-YOLO, a framework that unites a YOLO backbone, a lightweight transformer encoder, a gated cross-attention fusion module, and a Re-ID-aware multi-object tracker to deliver efficient, real-time performance in complex surveillance scenes. A learnable gating mechanism balances local and global features, and the framework attains 82.9% mAP@0.5, 77.3% MOTA, a 79-identity-switch reduction, and 42 FPS on an RTX 4090 across the MOT17 and MOT20 benchmarks, positioning HT-YOLO as a practical solution to the accuracy–latency trade-off.

Method, Experiments and Results

A. HT-YOLO System Architecture

The HT-YOLO pipeline follows a four-stage hierarchical architecture:
Stage 1 is to extract features. Using a YOLOv8-style CSP-Dark net backbone to extract multi-scale features maps with strides of 8, 16, 32, as P3, P4, and P5 respectively.
Stage 2: Global Context Modeling: In the case of long-range spatial relations, the deepest feature map is tokenized and processed by a compact transformer encoder.
Stage 3: Cross-Attention Fusion: A learnable, per-channel gated cross-attention module is used to fuse local and global representations.
Stage 4: Detection and Tracking: Kalman-filter and Re-ID tracking across frames is performed after multi-scale detection head regresses boxes and class scores from the fused features.

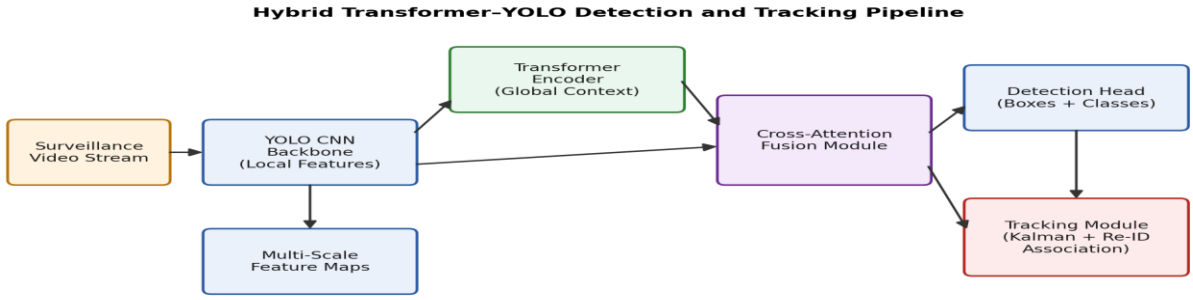

Fig. 1. Overview of the Hybrid Transformer–YOLO (HT-YOLO) Detection and Tracking Pipeline.

Figure 1 traces the pipeline end to end: the cross-attention fusion module merges global context tokens from the transformer encoder with local multi-scale features from the YOLO backbone, after which the detection head predicts boxes and classes that are passed to the tracking module for frame-to-frame identity association.

B. AI Model Stack

The model stack comprises four complementary components, following design principles from hybrid CNN-transformer architectures [4] and set-prediction transformers [2]:
Convolutional Backbone: CSP-Darknet for efficient local, multi-scale feature extraction.
Transformer Encoder: Four multi-head self-attention layers (eight heads each) over a spatially pooled token grid for global context.
Cross-Attention Fusion: A gated mechanism that adaptively balances local and global features per region.
Tracking Module: A Kalman filter with an appearance Re-ID embedding, associated via the Hungarian algorithm on a joint IoU–cosine-similarity cost matrix.

C. Module Details

Positional embeddings are added to the flattened P5 feature map before the transformer encoder processes it over a spatially pooled grid that reduces the token count fourfold, allowing each token to attend across the frame and link occluded object fragments to contextual cues such as local crowd density. Within the cross-attention fusion module, local features serve as queries and global tokens as keys and values, and a per-channel learnable gate weighs local detail against global context—favoring local detail where occlusion is low and global context where it is high. During tracking, a Kalman filter predicts each object's position from its motion history, and the Hungarian algorithm matches predicted boxes to new detections using a cost matrix that combines IoU with cosine similarity between Re-ID embeddings drawn from the fused features, which sustains identity through brief occlusions.

D. Loss Formulation

The network is trained end-to-end with a composite loss: a classification term, a Complete-IoU (CIoU) box-regression term, a distribution focal loss for boundary refinement, and a triplet-margin embedding loss encouraging compact intra-identity and separated inter-identity clusters. Classification and embedding losses are weighted 1:1, with CIoU regression weighted at 0.5 relative to classification; a gating regularisation term discourages the fusion gate from collapsing toward either branch.

E. Datasets and Experimental Setup

The framework is evaluated on MOT17 and MOT20—pedestrian-dense benchmarks with varying crowd density and occlusion—and a custom multi-camera surveillance dataset of approximately 120,000 annotated frames across twelve camera views under variable lighting. Training uses AdamW (initial learning rate 1e-4, cosine decay) for 300 epochs at 640×640 resolution on a single RTX 4090 GPU, with mosaic augmentation, random scaling, colour jittering, and horizontal flipping.

F. Prediction Performance Metrics

Table I benchmarks HT-YOLO against a YOLOv8 baseline and a DETR-style detector on the custom dataset. HT-YOLO reaches 82.9% mAP@0.5 and 77.3% MOTA—improvements of 4.6 and 6.1 points over YOLOv8—while holding 42 FPS, close to the YOLOv8 baseline and well ahead of the DETR-style detector's 19 FPS.

TABLE I. Detection and Tracking Performance on the Custom Surveillance Dataset.

Model	mAP@0.5 (%)	MOTA (%)	IDF1 (%)	FPS
YOLOv8 (baseline)	78.3	71.2	69.8	58
DETR-style detector	80.1	73.5	71.0	19
HT-YOLO (proposed)	82.9	77.3	75.4	42

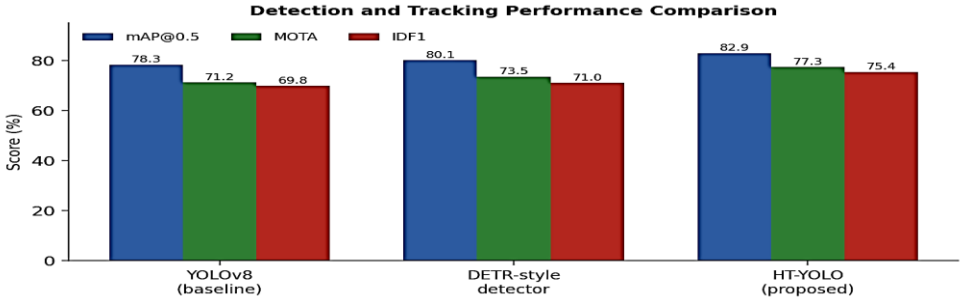

Fig. 2. Detection and Tracking Performance Comparison Across YOLOv8, DETR-style, and HT-YOLO Models.

Figure 2 compares mAP@0.5, MOTA, and IDF1 across the three models. HT-YOLO leads on every metric, with its widest relative margin in MOTA, suggesting that global context modeling benefits tracking stability even more than single-frame detection.

Across MOT17 and MOT20, HT-YOLO raises MOTA by 5.4 and 6.1 points, respectively, over a YOLOv8+DeepSORT baseline (Table II); the larger gain on MOT20, the more crowded of the two benchmarks, supports the view that global context matters most under heavy occlusion.

TABLE II. MOTA Comparison Across Evaluation Datasets.

Dataset	MOTA – YOLOv8 (%)	MOTA – HT-YOLO (%)	Δ (pp)
MOT17	68.9	74.3	+5.4
MOT20	60.2	66.3	+6.1
Custom (ours)	71.2	77.3	+6.1

G. Ablation Study

Table III confirms that both components contribute distinct benefits. Removing the transformer module costs 3.1 points of mAP@0.5, indicating that global context aids track continuity through occlusion. Replacing Re-ID association with IoU-only matching nearly doubles identity switches, from 142 to 231 (+62.7%).

TABLE III. Ablation Study on the Custom Surveillance Dataset.

Configuration	mAP@0.5 (%)	MOTA (%)	IDs
Full HT-YOLO	82.9	77.3	142
w/o Transformer module	79.8	74.0	168
w/o Re-ID (IoU-only)	82.1	72.6	231

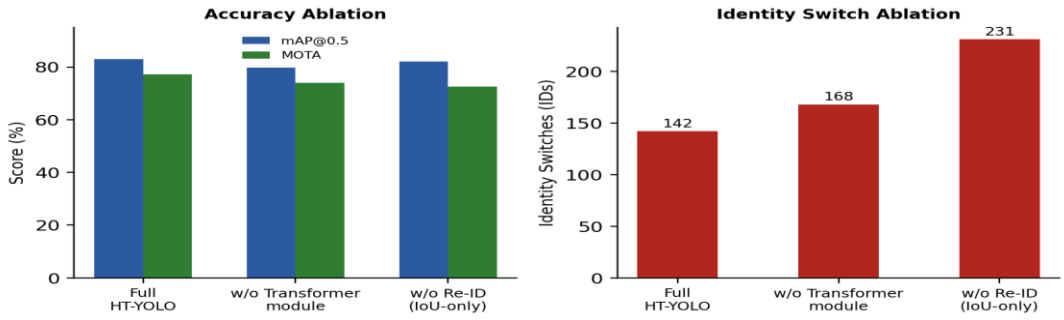

Fig. 3. Ablation Impact on Accuracy (Left) and Identity Switches (Right).

Figure 3 shows that the transformer module accounts for the larger drop in mAP@0.5 when removed, while the Re-ID embedding accounts for the larger rise in identity switches when removed, confirming that the two components address distinct failure modes.

H. Computational Efficiency and Limitations

HT-YOLO runs at 42 FPS at 640×640 on an RTX 4090, a modest drop from YOLOv8's 58 FPS but still well ahead of the DETR-style detector's 19 FPS and adequate for surveillance use. Three limitations remain: pooled tokens sacrifice some spatial precision, which slightly hurts small-object localization; the jointly trained Re-ID embedding may underperform a dedicated re-identification model under long-term

occlusion; and because results were obtained on a single high-end GPU, techniques such as attention pruning or knowledge distillation would be needed for edge deployment.

Discussions

HT-YOLO improves detection accuracy, tracking stability, and efficiency, yielding a more favorable accuracy–latency trade-off than convolution-only detectors. Relative to YOLOv8, it delivers a 4.6-point gain in mAP@0.5 and a 6.1-point gain in MOTA, with the largest benefit in occluded or distant regions, while a 16 FPS drop still leaves it within real-time range. The ablation study indicates that the transformer module chiefly improves detection accuracy under occlusion, while the Re-ID-aware tracker chiefly improves cross-frame identity stability. Both effects hold across MOT17, MOT20, and the custom dataset, and the largest absolute gain occurs on MOT20, the most congested benchmark, reinforcing the value of global context under heavy occlusion. Remaining limitations—a modest loss of small-object precision from token pooling and a reliance on a specific Re-ID training corpus—are natural directions for future work.

Conclusions

1. **Problem statement addressed / motivation:** Convolutional detectors have small receptive fields, which become a liability in surveillance video once occlusion and crowding set in. Transformer detectors address this limitation but are usually too computationally demanding for real-time deployment, motivating a hybrid design.
2. **Method used:** HT-YOLO combines a YOLO backbone, a lightweight transformer encoder for global context, a gated cross-attention fusion module, and a Kalman-filter/Re-ID tracker, trained end-to-end with a combined detection and embedding loss and evaluated on MOT17, MOT20, and a custom surveillance dataset.
3. **Key findings:** On the custom dataset, HT-YOLO gains 4.6 and 6.1 points in mAP@0.5 and MOTA over YOLOv8, and the ablation results show a consistent 5.4–6.1-point MOTA improvement across all three datasets while sustaining 42 FPS on an RTX 4090.
4. **Limitations and future work:** The jointly trained Re-ID embedding may be less effective than a dedicated model under long-term occlusion, and small-object precision is slightly reduced. Future work includes attention pruning and knowledge distillation for edge deployment, a dedicated Re-ID training corpus, multi-camera identity association, and temporal transformer modules for motion modeling.

References

1. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788.
2. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-End Object Detection with Transformers,” in Proc. European Conf. Computer Vision (ECCV), 2020, pp. 213–229.
3. X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable DETR: Deformable Transformers for End-to-End Object Detection,” in Proc. Int. Conf. Learning Representations (ICLR), 2021.
4. A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in Proc. Int. Conf. Learning Representations (ICLR), 2021.
5. C. Wang, A. Bochkovskiy, and H. Liao, “YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2023.
6. G. Jocher et al., “Ultralytics YOLOv8,” Ultralytics, 2023.
7. A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, “Simple Online and Realtime Tracking,” in Proc. IEEE Int. Conf. Image Processing (ICIP), 2016, pp. 3464–3468.

8. N. Wojke, A. Bewley, and D. Paulus, “Simple Online and Realtime Tracking with a Deep Association Metric,” in Proc. IEEE Int. Conf. Image Processing (ICIP), 2017, pp. 3645–3649.
9. A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, “MOT16: A Benchmark for Multi-Object Tracking,” arXiv:1603.00831, 2016.
10. P. Dendorfer et al., “MOT20: A Benchmark for Multi-Object Tracking in Crowded Scenes,” arXiv:2003.09003, 2020.