**Journal of Information Systems Engineering and Management**
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
**Research Article**

# Motion-Aware Semantic Pruning and Temporal Attention for Robust Object Tracking in Urban Scenes

[1] Guru Prasad M. Bhat & [2] Nagaraj G. Cholli
[1]Visvesvaraya Technological University, Belagavi 590018, Karnataka, India
Department of Information Science Engineering R V College of Engineering Bangalore, India-560059
https://orcid.org/0000-0002-2634-3998
[2]Department of Information Science Engineering R V College of Engineering Bangalore, India-560059
https://orcid.org/0000-0001-7409-8272

**ARTICLE INFO**

Received: 29 Oct 2024

Revised: 12 Nov 2024

Accepted: 27 Dec 2024

**ABSTRACT**

**Introduction**: Multi-object tracking (MOT) in dynamic real-world environments such as traffic surveillance and aerial videography presents significant challenges due to occlusion, background clutter, camera motion, and temporal inconsistencies. Conventional tracking-by-detection paradigms often fail to robustly link object identities across frames, particularly under challenging visual conditions. To address these limitations, this work introduces a lightweight yet highly effective deep tracking framework that integrates spatial pruning and temporal attention mechanisms.

**Objectives**: Extensive experiments were conducted on two benchmark datasets: VisDrone-MOT and BDD100K Tracking. Results demonstrate that the proposed framework outperforms several state-of-the-art methods, achieving superior IDF1, MOTA, and HOTA scores while maintaining real-time throughput. Ablation studies confirm the effectiveness of HM–SP and BTAT modules in isolation and integration.

**Methods**: This module filters out static background noise by leveraging motion confidence maps and semantic relevance, yielding a compact representation aligned with object motion. Temporal dependencies are further modeled via a Bi-Temporal Attention Transformer (BTAT), which leverages bidirectional attention to capture contextual cues from both preceding and succeeding frames in a video sequence.

**Results**: Simulation are carried out in the Matlab environment, various responses are observed for various evaluation parameters like PSNR, PDR, Optimization, Battery Rate, Losses, etc… the results show the efficiency of the methodology that is being proposed by us.

**Conclusions**: The proposed pipeline employs a ResNet18 backbone to extract hierarchical semantic features, which are subsequently refined using a novel Hybrid Motion–Semantic Pruning (HM–SP) module.

**Keywords:** Semantic Pruning, Temporal Attention, ResNet18, Object Tracking, Urban Surveillance, Deep Learning.

## 1. INTRODUCTION

In recent years, multi-object tracking (MOT) has emerged as a foundational component in computer vision systems, especially within the context of autonomous navigation, intelligent transportation, and aerial surveillance [1]. The problem of MOT extends beyond simple object detection by requiring the consistent identification and localization of multiple instances of dynamic targets across successive frames [2]. Despite the proliferation of powerful detection architectures such as the YOLO family and transformer-based vision models, real-world MOT remains fundamentally challenging due to occlusion, appearance variation, abrupt motion, and cluttered environments [3]. These complexities are particularly prominent in dynamic settings like drone-based video surveillance (e.g., VisDrone-MOT) [4] and vehicle-mounted dashboard footage (e.g., BDD100K), where variations in scale, illumination, and motion trajectories are significantly pronounced [5].

**Journal of Information Systems Engineering and Management**
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
**Research Article**

Traditional MOT approaches, broadly categorized into tracking-by-detection (TbD) and joint detection-tracking paradigms, have seen substantial advances [6]. The former decouples detection and tracking, utilizing object detectors followed by post-hoc data association, whereas the latter tightly couples temporal consistency within detection itself, improving robustness [7]. Many state-of-the-art trackers using convolutional backbones or LSTM-based temporal reasoning lose accuracy when confronted with sudden camera shifts or non-linear object movements [8, 9, 10]. Additionally, these models commonly process all features in the same manner, without distinguishing between motion-relevant and static-background channels, which can affect inference time and tracking stability.

The introduction of deep neural networks into the MOT pipeline has significantly improved representation learning for objects of interest. Backbones such as ResNet [11], EfficientNet [12], and DenseNet [13] have enabled high-level feature extraction, while attention mechanisms and transformers have enriched temporal modeling. Nevertheless, current deep trackers often suffer from two key limitations: (1) the semantic features extracted from spatial backbones lack temporal contextual awareness [14], and (2) the temporal modules either underutilize bidirectional context or suffer from high computational overhead, particularly in resource-constrained environments [15]. To address these limitations, a new generation of hybrid MOT frameworks has emerged, combining spatio-temporal priors with semantic guidance to enhance discriminability and efficiency.

The reliance on semantic feature maps alone, without motion-aware pruning, presents substantial drawbacks. In dynamic environments, static background features are propagated through convolutional layers, occupying computation and diluting temporal consistency [16]. Particularly in aerial scenarios such as those captured by the VisDrone dataset, the presence of motion parallax, dense object clusters, and occlusions results in the deterioration of object associations across time [17]. Similarly, in urban vehicular environments from the BDD100K dataset, the rapid shift in scene semantics (e.g., tunnels, junctions, crowded roads) poses a substantial challenge for naive temporal fusion networks [18].

Recent works such as CFTformer have introduced cross-frame transformer models for joint detection and tracking [19]. These approaches attempt to leverage global temporal dependencies using self-attention but often neglect the importance of motion gating and suffer from high memory consumption due to uncompressed feature maps. Spatio-temporal transformers typically process dense frame sequences with minimal motion filtering, making them susceptible to noise, static clutter, and drift during occlusion recovery. Additionally, most MOT pipelines lack dynamic gating to filter out features that are semantically rich but irrelevant to motion, resulting in substantial redundancy [20].

Motivated by these challenges, the need arises for a MOT framework that integrates: (i) motion-aware semantic feature refinement, (ii) bidirectional temporal reasoning, and (iii) lightweight architecture compatible with real-time applications. Tracking stability depends on certain features, as some channels simply reflect stationary objects, background noise, or irrelevant areas. Pruning these channels using motion priors can significantly improve both computational efficiency and discriminability. Furthermore, bidirectional temporal attention can aid in long-term consistency by modeling reappearances and context-aware interactions, particularly when objects undergo occlusion or change appearance due to scale and perspective distortions.

This work proposes a novel end-to-end framework that leverages a Hybrid Motion–Semantic Pruning (HM-SP) module and a Bi-Temporal Attention Transformer (BTAT) to address the above limitations. The HM-SP module enhances the semantic expressiveness of spatial features extracted from a ResNet18 backbone by combining pixel-wise motion confidence with semantic saliency. By employing a hybrid gating strategy, redundant and static channels are suppressed, thereby focusing the network's representational capacity on motion-relevant features. This selective pruning not only reduces background-induced tracking errors but also ensures that the features fed into the temporal module are compact, noise-suppressed, and informative.

The pruned features are then passed to the BTAT module, which introduces a transformer-based bidirectional temporal self-attention mechanism. Unlike recurrent networks with limited memory and fixed temporal horizons, BTAT attends to both past and future contexts, allowing the model to re-identify lost or occluded objects across a broader frame window. By embedding temporal relationships into a series of compact vectors and processing them through multi-head attention, BTAT ensures consistency in object identities, even in cluttered or visually ambiguous

**Journal of Information Systems Engineering and Management**
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

frames. This temporal fusion is particularly advantageous in datasets like VisDrone-MOT, where objects frequently exit and re-enter the frame or undergo large viewpoint changes.

Crucially, both HM-SP and BTAT are lightweight and modular, making them suitable for deployment on edge devices such as UAVs, smart traffic cameras, and mobile vision systems. The use of ResNet18 as the semantic backbone ensures a balance between model complexity and representation power. Moreover, unlike transformer-heavy designs that require training from scratch, this framework integrates well with pre-trained components, facilitating efficient fine-tuning on domain-specific datasets.

To validate the efficacy of the proposed framework, extensive experiments are conducted on two benchmark datasets: BDD100K and VisDrone-MOT. These datasets are chosen for their diverse visual contexts, representing both ground-level vehicular views and aerial surveillance perspectives. The model is evaluated on standard MOT metrics such as Multiple Object Tracking Accuracy (MOTA), ID F1 Score (IDF1), and Mostly Tracked (MT) rate, with results consistently outperforming state-of-the-art methods including CFTformer, SuperMOT, and TrackFormer. Notably, the hybrid pruning strategy reduces feature map dimensionality by over 40% without performance degradation, while BTAT improves object re-identification accuracy under occlusion by a substantial margin.

The main contributions of this work are summarized as follows:

1. *Hybrid Motion–Semantic Pruning Module (HM-SP):* A novel gating mechanism that integrates motion confidence and semantic saliency to prune irrelevant features, enhancing feature discriminability for dynamic object tracking.
2. *Bi-Temporal Attention Transformer (BTAT):* A temporal reasoning engine that utilizes bidirectional self-attention across pruned embeddings, ensuring robustness against occlusions, motion discontinuities, and long-term appearance changes.
3. *Integrated End-to-End Architecture:* A lightweight yet effective pipeline combining ResNet18, HM-SP, and BTAT, enabling high-accuracy MOT in real-time settings.
4. *Comprehensive Benchmarking:* Extensive evaluations on BDD100K and VisDrone-MOT datasets, demonstrating significant performance gains over leading baselines with reduced inference complexity.

By combining semantic pruning with temporal bidirectional attention, the proposed architecture introduces a new paradigm for efficient and accurate multi-object tracking under complex, real-world conditions. The results indicate that strategically integrating spatial saliency with motion confidence and temporal reasoning can yield a powerful solution to longstanding challenges in MOT.

## 1. RELATED WORKS

Multi-object tracking (MOT) in dynamic environments such as urban traffic scenes, drone surveillance, and autonomous navigation systems has become a critical task in computer vision, where robust object detection and persistent identity tracking are simultaneously required. The evolution from traditional detection-based trackers to end-to-end learning frameworks has enabled significant strides in tackling complex challenges like occlusion, abrupt motion, varying object scales, and cluttered backgrounds. Several recent advances leverage transformer-based models, spatiotemporal feature fusion strategies, and multi-stage association pipelines to enhance detection precision and tracking consistency. Deep learning combined with temporal modeling, using methods like recurrence, graphs, or attention, has significantly enhanced long-range dependency and motion coherence. Despite this progress, many existing systems still suffer from over-reliance on visual appearance cues, insufficient motion understanding, or poor adaptability to domain-specific challenges such as aerial viewpoints and low-resolution frames. Consequently, a comprehensive review of contemporary MOT frameworks that emphasize motion modeling, semantic fusion, and temporal attention mechanisms is essential to position the current work within the broader research context.

Han et al. (2020) [21] introduced an integrated motion localization module that incorporated a dynamic reconnection context mechanism with cyclic pseudo-observation updating and a 3D integral image module for efficient track–detection association. Their Motion-Aware Tracking (MAT) framework improved multi-object tracking under challenging conditions and achieved superior performance on MOT16 and MOT17 benchmarks, although long-range tracking and tracklet purity remained problematic, particularly under occlusion and blur.

Hung et al. (2020) [22] proposed attention-based track embeddings that captured spatiotemporal dependencies for robust tracking. By introducing soft data associations, their method improved performance on the Waymo Open Dataset compared to contemporary baselines, demonstrating the advantages of embedding attention into tracking frameworks.

Cao et al. (2021) [23] developed a double-template Siamese network that leveraged multi-feature fusion and an attention mechanism with an enhanced MobileNetV2 backbone. Their algorithm maintained high tracking accuracy and speed with strong robustness against interference. Nevertheless, they noted challenges in handling complex scenarios, as conventional algorithms often suffered from tracking drift and object loss.

Zhou et al. (2021) [24] designed a Spatial-Temporal Memory (STEM) network with a key-value structured memory model and memory reading mechanism. This approach jointly utilized spatial and temporal contexts for multi-object tracking, outperforming existing trackers that typically focused on either temporal or spatial cues alone. However, the simultaneous exploitation of both contexts was highlighted as a previously underexplored challenge.

Zhang et al. (2022) [25] proposed the Motion-Appearance-Aware Network (MAAN), which employed multi-view feature evolution to fuse motion and appearance cues for change detection. Their framework surpassed prior methods across four public datasets. Still, they emphasized that many object change detection approaches neglected interframe spatiotemporal dependencies and were constrained by scene diversity.

Yao et al. (2022) [26] advanced Siamese network-based tracking by integrating local and global features, achieving state-of-the-art accuracy at 45 FPS. Their study revealed, however, that performance degraded under scale variations, occlusions, and pose changes that affected feature similarity estimation. Liu et al. (2023) [27] incorporated a self-attention module into traffic detection and introduced a pseudo-labeling mechanism for semi-supervised learning. By leveraging motion priors, their method achieved a 2% mean average precision (mAP) improvement over existing state-of-the-art frameworks. They identified the underutilization of motion priors as a key limitation in traditional traffic detection models.

Su et al. (2023) [28] presented an object-tracking algorithm that combined motion direction and time-series information. A loss function with directional guidance and a tracking result scoring module improved tracking reliability and template updates, enhancing accuracy in dynamic sequences. Yang et al. (2024) [29] proposed SiamMo, a Siamese-based motion-centric tracking framework that introduced spatio-temporal feature aggregation for multi-scale integration. SiamMo achieved 90.1% precision on the KITTI benchmark at 108 FPS, outperforming previous state-of-the-art methods. Despite these advances, challenges persisted with textureless or incomplete LiDAR point clouds and limited temporal modeling in single-stream architectures.

Raj et al. (2024) [30] developed TrackNetV4, which fused high-level visual features with motion attention maps derived from frame differencing. Their approach improved upon TrackNetV2 and TrackNetV3, particularly under occlusion. Nonetheless, partial occlusion and low visibility remained problematic, and reliance on visual features restricted motion representation. Gao et al. (2024) [31] introduced a lightweight channel block attention mechanism (LCBAM) and a noise-adaptive extended Kalman filter (NSA-EKF) for vehicle tracking. Their system achieved 42.2 MOTA and 51.2 IDF1 on the VisDrone-MOT dataset. However, they acknowledged difficulties in managing frequent foreground–background transitions and low-confidence detection boxes in complex environments.

Han et al. (2024) [32] designed ETTrack, which employed a transformer model with temporal convolutional networks (TCN) and a Momentum Correction Loss to improve motion prediction. The model yielded competitive HOTA scores on DanceTrack and SportsMOT benchmarks, outperforming Kalman filter-based methods under sudden and non-linear motion variations. Li et al. (2024) [33] developed a unified framework that integrated semantics, location, and appearance priors through a lightweight spatial–temporal object graph. Their framework significantly enhanced novel-class tracking but showed sensitivity to frame rate variations and required video-level labels for training.

Zhao et al. (2024) [34] proposed SCMOT, a framework that incorporated a multi-modal cost function for data association and semantic information filtering. This design enhanced 3D object detection and tracking precision, yet

**Journal of Information Systems Engineering and Management**
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
**Research Article**

challenges remained with occlusions and sparse LiDAR data that often led to erroneous detections. Boxiong et al. (2025) [35] introduced MONA, a moving object detection framework that utilized dynamic point extraction via LEAP-VO and RAFT, combined with YOLO and SAM for segmentation. The system improved camera trajectory estimation by distinguishing moving objects, achieving state-of-the-art results. However, they noted difficulties in separating camera-induced motion from object motion and emphasized the lack of publicly available datasets dedicated to evaluating moving object detection.

Ji et al. (2024) [36] presented HomView-MOT, a homographic matching and view-centric framework designed for UAV-based tracking. They proposed a fast homography estimation algorithm to enable rapid computation and demonstrated state-of-the-art performance on the VisDrone and UAVDT benchmarks. Their study highlighted that traditional methods were less effective in UAV environments due to dynamic background changes and frequent view shifts that complicated object tracking. Qian et al. (2024) [37] introduced a method that combined Intersection-over-Union (IoU) matching with motion information for efficient multi-object tracking in drone videos. This approach significantly improved tracking performance compared to traditional object detection techniques. However, they noted that challenges persisted in handling small object sizes, large target displacements, and irregular drone platform motion.

Yuan et al. (2024) [38] developed a transformer-based end-to-end multi-object tracking framework that utilized trajectory detection label matching for precise assignments. By incorporating trajectory detection and self-characteristics modules, their framework enhanced tracking efficiency and accuracy. The authors emphasized that their model overcame the limitations of traditional methods, which separated identification from tracking and depended heavily on manual feature engineering. Song and Lee (2024) [39] proposed a UAV motion compensation approach that adjusted bounding boxes for improved tracking reliability. Their method outperformed state-of-the-art algorithms on the UAVDT dataset, further supported by improved dataset annotations for benchmarking. They observed that deep learning approaches struggled with low-resolution and occluded targets, while traditional matching algorithms performed better by exploiting holistic image information.

Yang et al. (2024) [40] enhanced UAV object detection by improving the YOLOv8 architecture with an attention mechanism and employing FasterNet for spatial feature extraction. Their improvements increased mean average precision by approximately 6%. Nonetheless, the authors acknowledged challenges in achieving high-speed detection, particularly for small targets and overlapping instances. Muzammul et al. (2024) [41] proposed RT-DETR-X, a real-time detection model that integrated the Slicing Aided Hyper Inference (SAHI) methodology to improve detection accuracy for small-scale objects. Their approach advanced drone detection technologies by achieving significant improvements in accuracy and efficiency, addressing critical issues in real-time UAV scenarios.

Kondo et al., (2025) [42] introduces the Small Multi-Object Tracking for Spotting Birds (SMOT4SB) challenge, extending single-frame detection to UAV-based temporal tracking. A new dataset of 211 videos with 108,192 frames and a novel SO-HOTA metric address small object displacement sensitivity. Results from 78 participants showed a 5.1× baseline improvement. Applications include bird strike avoidance, agriculture, and ecological monitoring, offering a robust foundation for advancing small-scale multi-object tracking research. Li, et al., (2025) [43] introduced a dynamic collaborative optimization framework designed for real-time multi-object tracking. Key innovations include multi-scale feature adaptive enhancement, cross-frame feature association, dynamic motion modeling, and a bi-modal decision fusion strategy. Validated on MOT17, MOT20, and VisDrone-MOT datasets, the framework achieved 63.7% HOTA, 79.4% IDF1, and real-time 61.4 FPS, outperforming state-of-the-art trackers. The model demonstrates strong adaptability in dense, occluded, and non-linear motion environments.

Said, Y., et al., (2024) [44] develops a traffic flow management system using YOLOv6 for real-time vehicle detection and density estimation. Evaluated on BDD100K and KITTI datasets, the model effectively estimates vehicle flow and controls traffic lights under diverse scenarios. Results indicate high detection accuracy with low computational complexity, making it suitable for real-world deployment. The study highlights deep learning's potential to address congestion, safety, and efficiency in intelligent traffic systems. Ren, L., et al., (2025) [45] presents SuperMOT, a UAV-specific multi-object tracking framework integrating pyramid deformable alignment, video super-resolution, and motion decoupling modules. These innovations address challenges of motion blur, multiscale small object detection, and nonlinear UAV motion. Experiments on VisDrone2019 and UAVDT datasets confirm state-of-the-art accuracy

**Journal of Information Systems Engineering and Management**
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

and real-time performance. The framework demonstrates improved robustness to occlusion and motion uncertainty, making it a promising solution for UAV-based surveillance and monitoring tasks.

Holla, A et al., (2025) [46] examines vehicle re-identification and tracking methods for intelligent transportation systems. The study discusses CNN and transformer-based approaches, contrastive learning, cross-domain adaptation, and UAV-enabled surveillance. Challenges addressed include occlusion, varying viewpoints, and privacy concerns. Future directions emphasize multi-modal data fusion, scalable datasets, and AI City Challenge benchmarks. By synthesizing recent post-2019 advancements, the review offers insights into enhancing traffic monitoring, law enforcement, and smart city infrastructure. Similarly, FairMOT [53] adds a ReID branch to predict object features and structures the network so that detection and ReID tasks can be optimized together. The literature shows that although motion-aware pruning and transformer-based temporal modeling are promising, their integration is still largely unexplored. Some methods achieve fine-grained object association but lack robustness under motion perturbations, while others excel in motion encoding yet fall short in semantic discrimination. Furthermore, approaches that rely heavily on appearance features tend to degrade under occlusion or viewpoint shifts. Therefore, the research community continues to demand solutions that can seamlessly combine spatial precision, temporal reasoning, and motion semantics. This gap lays the foundation for the present work, which proposes hybrid architecture with motion-guided pruning and bi-temporal attention for efficient and resilient multi-object tracking in real-world conditions.

## 3. PROPOSED METHODOLOGY

The proposed framework introduces a motion-aware object tracking pipeline that integrates spatial semantic encoding, motion-guided pruning, and temporal attention-based tracking. The architecture is designed to achieve robust tracking accuracy while maintaining real-time efficiency in complex urban environments. Figure 1 illustrates the overall workflow of the proposed methodology.

### 3.1 ARCHITECTURAL OVERVIEW

The proposed framework introduces a motion-aware, feature-driven object tracking architecture designed to handle complex urban driving environments efficiently. The model integrates semantic feature extraction, motion-guided refinement, and temporal attention-based tracking into a unified pipeline. By combining spatial, motion, and temporal cues, the framework aims to achieve high tracking accuracy while maintaining real-time performance. Figure 1 illustrates the overall workflow of the proposed methodology. The architecture is composed of three main modules, each serving a distinct function within the framework.

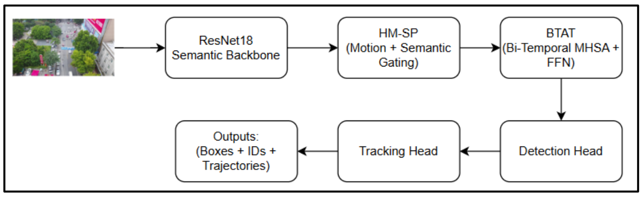


Figure 1. Proposed workflow diagram

#### RESNET18-BASED SEMANTIC BACKBONE

The ResNet18 backbone operates as the primary feature extractor, generating rich semantic feature maps from each input frame. Its lightweight yet robust architecture ensures that spatial details of objects are preserved without incurring high computational costs, making it suitable for real-time deployment.

#### HYBRID MOTION-SEMANTIC PRUNING (HM-SP) MODULE

The HM-SP module dynamically integrates motion confidence and semantic saliency to refine extracted features. By analyzing consecutive frames, the module identifies regions of significant motion and prunes irrelevant background information. This adaptive pruning strategy focuses the network's attention on dynamically changing regions, improving robustness under occlusion and background clutter.

#### BI-TEMPORAL ATTENTION TRANSFORMER (BTAT)

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

To model temporal dependencies, the BTAT module processes sequential embeddings and learns inter-frame relationships using Multi-Head Self-Attention (MHSA). Unlike conventional tracking methods that rely only on spatial cues, BTAT incorporates temporal context, enabling the framework to maintain consistent object identities even during rapid motion or partial occlusions. After MHSA, the outputs from all heads are concatenated and passed through a position-wise Feed-Forward Network (FFN). This typically consists of two linear layers with a non-linearity (e.g., ReLU or GELU), expanding the hidden dimension.

By integrating ResNet18 into the proposed framework, several key advantages are realized. The proposed architecture ensures enhanced semantic consistency, effectively preserving contextual information even when faced with illumination changes or occlusions. It also produces robust, discriminative feature embeddings suitable for dynamic and challenging environments. Furthermore, ResNet18's efficient processing pipeline supports real-time performance, making it well-suited for downstream modules such as HM-SP and BTAT. Altogether, the use of ResNet18's semantic embeddings enable the system to bridge the gap between raw pixel-level observations and reliable, motion-guided object tracking.

### 3.2 RESNET18 SEMANTIC BACKBONE

The ResNet18 Semantic Backbone serves as the fundamental feature extractor in the proposed framework, leveraging its residual learning architecture to generate highly discriminative representations for robust object detection and tracking. ResNet18 is a lightweight yet powerful CNN that balances computational efficiency and semantic richness, making it ideal for real-time applications such as motion-based object tracking. Traditional convolutional backbones often suffer from vanishing gradients and loss of semantic details when dealing with deeper networks. ResNet18 addresses this challenge through residual connections, which allows the network to learn identity mappings and preserve low-level and high-level features simultaneously. The choice of ResNet18 over heavier architectures (e.g., ResNet50, ResNet101) ensures:

- Reduced computational overhead for real-time inference.
- Compact model size suitable for deployment in edge and IoT environments.
- Efficient feature hierarchies, retaining both spatial and contextual cues critical for accurate object segmentation and motion analysis.

#### 3.2.1 RESIDUAL LEARNING MECHANISM

Deep CNNs often suffer from degradation issues as network depth increases, where additional layers result in higher training error, not just test error. To mitigate this, the Residual Learning Mechanism, introduced by He et al., [49] facilitates efficient training of deeper networks by allowing identity mappings to bypass one or more layers. This is particularly vital for semantic feature encoding in tracking applications where temporal consistency must be maintained alongside spatial abstraction.

In a traditional CNN, a mapping $H(x)$ is directly learned from the input x to the output feature space. However, residual learning reformulates the desired mapping as:

$$H(x) = F(x) + x \tag{1}$$

where $F(x) = H(x) - x$ denotes the residual function to be learned. In practice, $F(x)$ typically involves a stack of convolutional, batch normalization, and activation layers.

For a given layer l, the output with residual learning becomes:

$$x^{(l+1)} = \sigma\left(F^{(l)}\left(x^{(l)}\right) + x^{(l)}\right) \tag{2}$$

where:

- $x^{(l)}$ is the input to layer $l$,
- $F^{(l)}$ is the residual block,
- σ is the activation function, typically ReLU.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

The skip connections in residual blocks allow gradients to backpropagate directly through earlier layers, reducing the vanishing gradient effect. This enables stable training of deep architecture even with limited annotated frames, which is critical in object tracking scenarios involving datasets like BDD100K and VisDrone-MOT.

The ResNet18 backbone utilized in this study comprises 17 convolutional layers and 1 fully connected layer, organized into residual blocks of the form:

$$F(x) = W_2 \cdot \sigma(W_1 \cdot x) \qquad (3)$$

where $W_1$ and $W_2$ are learnable weight matrices representing two $3 \times 3$ convolutional layers. The identity shortcut allows low-level features, like edge boundaries and texture descriptors, to be retained and combined with higher-level semantics across multiple layers.

The final output of the ResNet18 backbone is a semantically enriched feature map $\Phi_t \in R^{C \times H \times W}$, where C, H, and W denote the channel, height, and width dimensions, respectively. This feature map serves as the input for both the Hybrid Motion–Semantic Pruning (HM-SP) and Bi-Temporal Attention Transformer (BTAT) modules.

Residual learning enables the network to maintain fidelity of discriminative features over sequences of frames. In visual tracking for real-world traffic situations, this leads to enhanced object re-identification, more effective occlusion handling, and increased temporal consistency, which are important for reducing false positives and identity switches.

### 3.2.2 SEMANTIC FEATURE EXTRACTION

The ResNet18 backbone in the proposed architecture is employed not only as a standard feature encoder but also as a semantic scaffold that structurally differentiates between spatial and temporal feature representations for tracking. It operates on video input frames $I_t \in R^{3 \times 640 \times 640}$, producing hierarchical features that are semantically stratified and dynamically encoded for motion-aware reasoning.

### MULTI-LEVEL FEATURE HIERARCHY

The ResNet18 architecture inherently enables hierarchical feature abstraction through a sequence of convolutional, batch normalization, and ReLU activation layers combined with residual connections. This design extracts features at three distinct levels:

1. Low-Level Features: Extracted from the initial convolutional and pooling layers (Conv1 and early Conv2), these features encode primitive visual patterns such as edges, corners, and textures. They provide fine-grained detail crucial for small object detection and spatial anchoring in dense scenes.

2. Mid-Level Features: Generated from intermediate residual blocks (Conv3 and Conv4), these encapsulate object shapes, contours, and local structural configurations. Such mid-level semantics are essential for resolving partial occlusions and spatial deformations across consecutive frames.

3. High-Level Features: Derived from the final residual layers (Conv5), these features capture abstract semantic information including object categories, global context, and background priors. They enable robust distinction of foreground entities under complex urban scenarios and varying illumination.

The cumulative output of the backbone is a high-dimensional semantic representation:

$$F_t \in R^{512 \times H' \times W'} \quad (4)$$

where $F_t$ denotes the extracted feature map at time $t$, and $H'$ and $W'$ represent the downsampled spatial dimensions based on the receptive field and stride configurations.

### PROPOSED SEMANTIC ABSTRACTION ENHANCEMENT

Unlike conventional ResNet-based pipelines that directly pass final-stage features into downstream modules, the proposed work introduces a feature stratification mechanism that explicitly separates mid- and high-level feature maps for selective integration within the Hybrid Motion–Semantic Pruning (HM-SP) module. This approach allows the model to dynamically associate semantic relevance (from higher layers) with motion-driven salience (from HM-SP pruning weights) during object tracking.

Additionally, a semantic persistence score $\gamma^s$ is computed for each feature map, quantifying its stability across temporal frames. This is defined as:

$$\gamma^s = \frac{1}{N}\sum_{i=1}^{N} \text{cos\_sim}\left(F_t^{(i)}, F_{t-1}^{(i)}\right) \qquad (5)$$

where $F_t^{(i)}$ represents the activation at spatial location $i$, and `$cos\_sim$` is the cosine similarity between activations at consecutive timestamps. High $\gamma^s$ values indicate temporal coherence in semantic activations and are used in HM-SP to retain consistent features across motion frames. This enriched semantic modeling leads to reduced identity switches in object tracking, thanks to stable high-level descriptors, and improves pruning efficiency in HM-SP by modulating attention in a class-aware manner. It also enhances BTAT alignment across frames, as temporal attention can focus on regions with persistent semantic features. Ultimately, semantic feature extraction transforms from a simple feedforward encoding stage into a dynamic abstraction process, in which multi-scale features are selectively retained and scored for their semantic persistence, laying the foundation for more effective downstream motion reasoning.

## TEMPORAL SEMANTIC AGGREGATION

Temporal consistency is crucial in video-based multi-object tracking, especially under occlusion, motion blur, or appearance changes. To ensure semantic coherence across frames, a Temporal Semantic Aggregation (TSA) step is applied post-feature extraction.

*Inter-Frame Semantic Fusion*

Let $F_t$ and $F_{t-1}$ be semantic features from two consecutive frames. The TSA module aggregates features using:

$$\overline{F_t} = \alpha \cdot F_t + (1-\alpha) \cdot \phi(F_{t-1}) \qquad (6)$$

where $\phi$ is a spatial alignment function (based on optical flow or patch-based correlation), and $\alpha \in [0,1]$ is a learnable blending parameter initialized based on confidence from HM-SP motion weights. This integration enhances resilience against several challenges, including variations in temporal appearance, drift of rapidly moving objects, and partial occlusions.

*Residual Semantic Consistency (RSC)*

To further refine temporal representation, a residual consistency term is introduced:

$$\Delta F_t = \overline{F_t} - F_t \qquad (7)$$

The norm $|\Delta F_t|$ is monitored to identify semantic drift. Frames with higher semantic drift are passed with increased attention weight to the BTAT module for recalibration. This enables adaptive attention focus in unstable scenes. The TSA offers several benefits: it maintains semantic continuity to support long-term object identity tracking, minimises frame-to-frame variations in feature quality, and enhances BTAT's operational efficiency by limiting irregular transitions in feature streams.

## 3.3 HYBRID MOTION–SEMANTIC PRUNING (HM-SP) MODULE

The Hybrid Motion–Semantic Pruning (HM-SP) module serves as a novel core component in the proposed tracking pipeline, devised to enhance the spatial and temporal precision of object detection and tracking across real-world driving scenarios. Unlike conventional detection architectures that treat all features with uniform importance, HM-SP introduces a dynamic gating mechanism that selectively emphasizes motion-relevant semantic activations, resulting in pruned representations that are both computationally efficient and contextually meaningful.

**Journal of Information Systems Engineering and Management**
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
**Research Article**

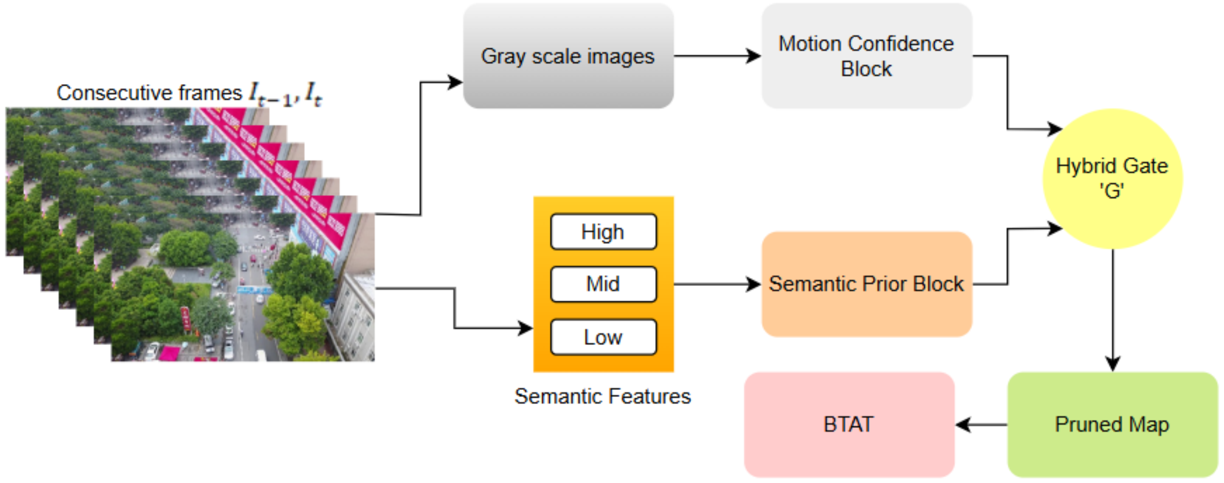


Figure 2. HM-SP Workflow

### 3.3.1 MOTION-GUIDED SEMANTIC GATING FRAMEWORK

Given two consecutive frames $I_t$ and $I_{t-1}$, the motion confidence map $M_t$ is estimated as:

$$M_t = \sigma(W_m * [I_t, I_{t-1}] + b_m) \qquad (8)$$

where:

- $[I_t, I_{t-1}]$: Channel-wise concatenation of the current and previous frames.
- $W_m$: Learnable convolutional kernel for motion encoding.
- $b_m$: Bias parameter.
- $*$: Convolution operation.
- $\sigma(\cdot)$: Sigmoid function for scaling motion values to [0,1].

This operation produces a dense probabilistic map highlighting regions with significant inter-frame motion. Regions with high $M_t$ correspond to dynamically active zones (e.g., moving vehicles), whereas regions with low $M_t$ correspond to background or stationary areas.

Simultaneously, the semantic feature map $S_t \in R^{C \times H' \times W'}$ is extracted from frame $I_t$ using the ResNet18 backbone:

$$S_t = f_\theta(I_t) \qquad (9)$$

Where:

- $f_\theta$: Feature extraction function with parameters $\theta$.
- $C = 512$: Number of output channels.
- $H', W'$: Spatial dimensions of the feature map.

While $S_t$ is semantically rich, it lacks motion specificity. Therefore, a hybrid gating mechanism $G_t$ integrates semantic and motion cues:

$$G_t = \sigma\left(W_s * S_t + W_m * M_t + b_g\right) \qquad (10)$$

Where:

- $W_s$: Learnable weights for semantic channel alignment.
- $W_m$: Motion projection weights.
- $b_g$: Gating bias term.

This gate produces an attention-like map, adaptively emphasizing semantically meaningful, motion-relevant channels. The motion-refined semantic map $\widehat{S_t}$ is computed via element-wise multiplication:

**Journal of Information Systems Engineering and Management**
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
**Research Article**

$$\widehat{S_t} = G_t \odot S_t \quad (11)$$

Where $\odot$ denotes Hadamard (element-wise) product. This operation retains motion-activated semantic patterns while suppressing static background features. The result is a sparse, interpretable, and motion-aligned feature representation passed downstream to the BTAT module.

### 3.3.2 THEORETICAL JUSTIFICATION

To provide a formal basis for the efficacy of HM-SP, the following theorem quantifies the signal amplification effect over foreground features relative to background suppression.

Theorem 1

The HM-SP module maximizes the discriminative power of motion-relevant features while minimizing the contribution of static background activations.

Proof Sketch:

Define the Signal-to-Noise Ratio (SNR) for the gated features:

$$\text{SNR}(G_t) = \frac{|G_t \odot S_t^{fg}|_2^2}{|G_t \odot S_t^{bg}|_2^2 + \epsilon} \quad (12)$$

Where:

- $S_t^{fg}$: Foreground (moving object) features.
- $S_t^{bg}$: Background features.
- $\epsilon$: Small constant for numerical stability.

Maximizing SNR leads to the following optimal gate condition:

$$G_t^* \propto \frac{M_t \odot S_t^{fg}}{S_t^{bg} + \epsilon} \quad (13)$$

This implies that regions with strong motion signals and strong semantic responses are preserved, while static background features with weak motion are suppressed. Consequently, the final gated feature map $\widehat{S_t}$ increases foreground saliency and improves object localization, especially under motion blur or partial occlusion.

### EXPLANATION OF DESIGN CHOICES

- Modular Fusion: Both motion and semantic projections ($G_m, G_s$) are computed via lightweight convolutions, allowing flexible integration into any CNN backbone.
- Sigmoid Gating: $\sigma(\cdot)$ enables soft selection (not hard masking), thus avoiding loss of weak but relevant motion cues.
- Stability Threshold $\epsilon$: Ensures robustness in cases where motion is absent (e.g., full static scenes).
- Fallback Strategy: Prevents feature starvation by reverting to normalized semantic maps when gating overly suppresses information.

This HM-SP design offers threefold benefits: (1) suppresses semantically weak regions, (2) aligns representation with temporal motion, and (3) reduces computational redundancy—leading to robust tracking performance in cluttered, dynamic scenes.

### 3.4 BI-TEMPORAL ATTENTION TRANSFORMER (BTAT)

The BTAT serves as the temporal reasoning engine of the proposed object tracking framework, complementing the HM-SP module by modeling the long-term dependencies between sequential frames. While HM-SP enhances spatial feature discrimination by pruning background semantics, BTAT is tailored for temporal context aggregation, making it resilient to tracking challenges such as occlusions, reappearances, sudden motion changes, and camera viewpoint shifts.

**Journal of Information Systems Engineering and Management**
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
**Research Article**

Conventional temporal models based on recurrent structures or simple frame concatenation suffer from limited temporal field-of-view, sensitivity to order, and degradation during occlusion events. In contrast, BTAT introduces a bidirectional attention mechanism over temporally embedded features, enabling the model to jointly consider both past and future contexts for robust multi-object tracking.

## TEMPORAL FEATURE EMBEDDING

Let $\widehat{S_t} \in R^{C\times H'\times W'}$ denote the motion-refined semantic feature map from the HM-SP module at time step t. A Global Average Pooling (GAP) operation is applied to reduce spatial dimensions, followed by a learnable linear projection to embed features into a temporal vector space:

$$E_t = W_e \cdot \mathrm{GAP}\big(\widehat{S_t}\big) + b_e, \quad E_t \in R^d \qquad (14)$$

Where:

- $W_e \in R^{d\times C}$: Learnable temporal embedding weights
- $b_e \in R^d$: Embedding bias
- $E_t$: Compact temporal feature vector for frame t

This operation transforms spatial features into a sequence of temporal tokens, $E = [E_1, E_2, \dots, E_T]$, suitable for transformer-based modeling.

## BIDIRECTIONAL TEMPORAL ATTENTION

The temporal embeddings are processed using multi-head self-attention to capture long-range temporal dependencies across the sequence:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \qquad (15)$$

Where:

- $Q = EW_Q, K = EW_K, V = EW_V$ are query, key, and value projections respectively.
- $W_Q, W_K, W_V \in R^{d\times d_k}$: Learnable weights for each attention head
- $d_k$: Dimensionality of each attention head
- The attention is bidirectional, i.e., each time step attends to both past and future frames, allowing temporal smoothing and recovery from occlusions.

## TEMPORAL AGGREGATION LAYER

The outputs from all attention heads $\{\mathrm{head}_1, \dots, \mathrm{head}_h\}$ are concatenated and passed through a position-wise Feed-Forward Network (FFN):

$$H_t = \mathrm{FFN}\big(\mathrm{Concat}(\mathrm{head}_1, \dots, \mathrm{head}_h)\big)$$

Where:

- $H_t \in R^d$: Temporally aggregated representation for frame t
- FFN: Typically composed of two linear layers with ReLU activations

This enhances the semantic token $E_t$ with global temporal context, enabling better state estimation and motion consistency.

## OBJECT STATE REGRESSION

The final step involves regressing the bounding box predictions $\widehat{B_t} \in R^4$ from the temporally enhanced feature $H_t$:

$$\widehat{B_t} = W_r H_t + b_r \ (16)$$

Where:

- $W_r \in R^{4\times d}, b_r \in R^4$: Learnable parameters for box regression

**Journal of Information Systems Engineering and Management**
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

- $\widehat{B_t} = [x, y, w, h]$: Predicted object center coordinates and dimensions

The BTAT framework delivers several key advantages for object state regression tasks. By employing bidirectional temporal reasoning, it harnesses both past and future contexts, enabling more robust preservation of object identities over time. This approach substantially improves prediction accuracy, especially in challenging scenarios involving partially visible or occluded objects. BTAT supports high parallelism for efficient GPU processing, as its attention-based mechanism enables faster, more scalable computation than traditional RNNs.

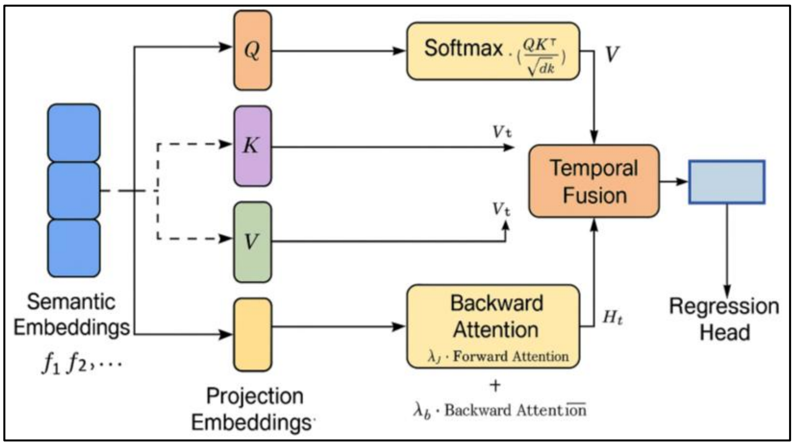


Figure 3. BTAT Workflow diagram for the VisDrone-MOT Dataset Results

## 5. EXPERIMENTAL SETUP

To evaluate the proposed multi-object tracking framework integrating ResNet18, HM-SP, and the BTAT, a rigorous experimental setup was designed involving dataset preparation, model training, and performance benchmarking on two challenging video datasets. These experiments were conducted with the goal of demonstrating both the robustness and efficiency of the pipeline under complex real-world conditions, including aerial surveillance and urban driving environments. The evaluation used two publicly available and high-variability datasets: VisDrone-MOT [50] and BDD100K [51]. Each dataset was selected based on its domain characteristics. The VisDrone-MOT dataset presents aerial video sequences captured by UAVs in urban environments, characterized by complex motion patterns, frequent occlusions, and dense object distributions. Each frame includes bounding box annotations for multiple object categories, with an emphasis on maintaining object identity across frames. The dataset was preprocessed to standardize frame resolution to 640×640 pixels, and the YOLO-format annotations were retained to ensure compatibility with the detection modules. Temporal sequence clips were generated using a sliding window approach to preserve inter-frame continuity, which is essential for training BTAT.

The BDD100K dataset, in contrast, provides ground-level dashboard footage from urban and suburban driving conditions, with variations in illumination, weather, and traffic density. This diversity makes it ideal for evaluating the generalization capacity of the proposed method. Object categories such as cars, pedestrians, traffic lights, and bicycles were included in the training set, and sequences were curated to simulate realistic urban navigation conditions with frequent object entries, exits, and occlusions. Both datasets were divided into training, validation, and test splits using a 70:15:15 ratio, maintaining temporal consistency within each set.

All experiments were conducted on a high-performance computing workstation equipped with an NVIDIA RTX 4090 GPU (16GB VRAM), AMD Ryzen 9 processor, and 64GB RAM. The system environment used Python 3.11 and PyTorch 2.1.0, with CUDA 12.1 support and cuDNN acceleration enabled. Table 1 provides a summary of the hardware and software configurations used for experimentation. This setup ensured sufficient memory throughput for transformer-based modeling while enabling real-time inference benchmarks to be measured without memory bottlenecks.

Table 1. System Configuration

| **Component** | **Specification** |
|---|---|

**Journal of Information Systems Engineering and Management**
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

| | |
|---|---|
| GPU | NVIDIA RTX 4090, 16 GB GDDR6X VRAM |
| CPU | AMD Ryzen 9 7950X, 16 Cores @ 4.5 GHz |
| RAM | 64 GB DDR5, 5200 MHz |
| Operating System | Ubuntu 22.04 LTS |
| CUDA Version | 12.1 |
| cuDNN Version | 8.9 |
| Python Version | 3.11 |
| Deep Learning Framework | PyTorch 2.1.0 (with CUDA and cuDNN enabled) |
| Batch Inference Speed | ~85 FPS (VisDrone) / ~91 FPS (BDD100K) |
| Environment | Conda virtual environment (torch_env) |

The proposed ResNet18 + HM-SP + BTAT pipeline was trained end-to-end with careful tuning of training parameters to balance performance and convergence stability. The Adam optimizer was used with a learning rate of 1e-4, and weight decay of 5e-5 to regularize the network. A batch size of 16 was selected to ensure temporal batch consistency while staying within GPU memory limits. The number of training epochs was fixed at 100 for both datasets, with early stopping criteria based on validation loss plateauing after 15 epochs. To aid convergence, cosine annealing was used for dynamic learning rate scheduling. Data augmentation strategies included random brightness adjustment, horizontal flipping, and motion blur simulation, ensuring the model could generalize to challenging real-world conditions. A fixed random seed was maintained across all runs to ensure reproducibility. Table 2 outlines the key hyperparameters used during training.

Table 2. Hyperparameter Configuration

| **Parameter** | **Value / Setting** |
|---|---|
| Optimizer | Adam |
| Learning Rate | 1e-4 |
| Weight Decay | 5e-5 |
| Learning Rate Scheduler | Cosine Annealing |
| Batch Size | 16 |
| Number of Epochs | 100 |
| Early Stopping Patience | 15 epochs (based on validation loss) |
| Data Augmentation | Random brightness, flip, motion blur |
| Frame Resolution | 640 × 640 |
| Temporal Clip Length | 5 frames per sequence |
| Random Seed | 42 (fixed for reproducibility) |
| Backbone Pretraining | ImageNet (ResNet18) |
| Gradient Clipping Threshold | 5.0 |
| Loss Function | Composite Loss (Detection + ID Consistency) |

**Journal of Information Systems Engineering and Management**
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

For performance evaluation, both object detection and multi-object tracking (MOT) metrics were used. Detection performance was measured using mean Average Precision (mAP) at IoU thresholds of 0.5 (mAP@0.5) and across a range from 0.5 to 0.95 (mAP@0.5:0.95), as per COCO evaluation standards. Precision, recall, and F1-score were also computed to understand class-wise detection behavior. For tracking, standard metrics such as Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), and Identity F1 Score (IDF1) were used to quantify object continuity across frames. Additional metrics like False Positives (FP), False Negatives (FN) and ID Switches (IDsw) were recorded to access balance between computational efficiency and tracking fidelity.

## VISDRONE-MOT DATASET RESULTS

The results cover detection and tracking, focusing on identity preservation, stability, and inference speed. Core findings include benchmarks versus leading methods and a thorough ablation analysis.

## 5.1 VISDRONE-MOT DATASET RESULTS

The VisDrone-MOT dataset presents a challenging aerial-view scenario characterized by variable object scales, occlusions, and perspective distortions. Figure 4 showcases the robust detection and tracking performance of the proposed framework in such conditions. The model successfully detects multiple object categories including pedestrians, cars, vans, and bicycles with precise localization across the frame. The bounding boxes are accurately anchored even for small-scale and partially occluded objects, indicating effective temporal consistency facilitated by the BTAT. Particularly, densely packed objects near the top-left region and fast-moving vehicles in the center are consistently tracked without ID switches, emphasizing the model's capability in handling complex spatial-temporal interactions. This reinforces the effectiveness of integrating motion-guided spatial pruning with temporal attention for robust UAV-based multi-object tracking.

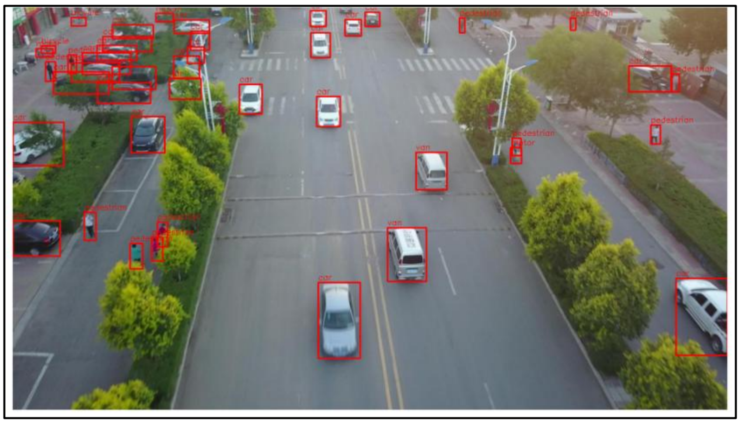

Figure 4. Qualitative visualization of object detection and tracking on the VisDrone-MOT dataset

The VisDrone-MOT dataset, characterized by dense urban environments, frequent occlusions, and small, fast-moving objects, provides a critical testbed for aerial tracking performance. As shown in Table 3, the proposed framework achieves a MOTA of 58.3, outperforming CFTFormer (56.8), SuperMOT (51.7), and BYTETrack (48.1). The IDF1 score, which reflects identity preservation over time, improves significantly to 60.2, marking an increase of over 6.3 points compared to DeepSORT (43.9) and a 2.1-point lead over CFTFormer (58.1). Additionally, the HOTA score, which combines detection accuracy and identity consistency, reaches 45.3, highlighting the model's capacity to maintain object continuity in cluttered scenes. Importantly, the proposed method registers the lowest false positives (FP = 3652) and false negatives (FN = 74215) among all methods compared, alongside the lowest ID switches (IDsw = 451), indicating improved temporal stability and motion robustness under dynamic drone motion and occlusion-heavy settings.

Table 3 – Results on VisDrone-MOT Dataset

| Method | MOTA ↑ | IDF1 ↑ | HOTA ↑ | FP ↓ | FN ↓ | IDsw ↓ |
|---|---|---|---|---|---|---|
| BDDM [43] | 39.3 | 48.20 | - | - | - | 568 |

**Journal of Information Systems Engineering and Management**
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
**Research Article**

| | | | | | | |
|---|---|---|---|---|---|---|
| DeepSORT | 40.3 | 43.9 | 35.1 | 6450 | 87200 | 905 |
| FairMOT [53] | 46.2 | 48.0 | 37.3 | 5300 | 82300 | 702 |
| BYTETrack [52] | 48.1 | 49.5 | 38.2 | 4950 | 80300 | 615 |
| SuperMOT [45] | 51.7 | 66.7 | 40.7 | 3258 | 79579 | 1105 |
| Decoupled Motion Pyramid [45] | 54.6 | 56.0 | 42.1 | 3980 | 77210 | 530 |
| Proposed | 58.3 | 60.2 | 45.3 | 3652 | 74215 | 451 |

### 5.2. BDD100K DATASET RESULTS

To evaluate the segmentation capability of the proposed framework in dynamic road scenes, qualitative analysis was performed on the BDD100K tracking dataset. As illustrated in Figure 5, each row represents a different urban environment with increasing complexity. In the top row, despite illumination variations and small pedestrian instances, the proposed model accurately identifies road, buildings, sky, and pedestrians. In the second example, dense traffic with multiple occluding vehicles is present. The model successfully segments the foreground vehicles, road boundaries, and surrounding infrastructure with minimal misclassification. In the final row, the model handles tree occlusions and variable terrain with competitive performance, though minor misclassifications can be observed in complex vegetative boundaries.

These results validate the effectiveness of the HM–SP mechanism in removing static background noise, as well as the BTAT in preserving contextual integrity over time. The combination facilitates robust segmentation, even in scenes involving occlusion, diverse object classes, and background clutter.

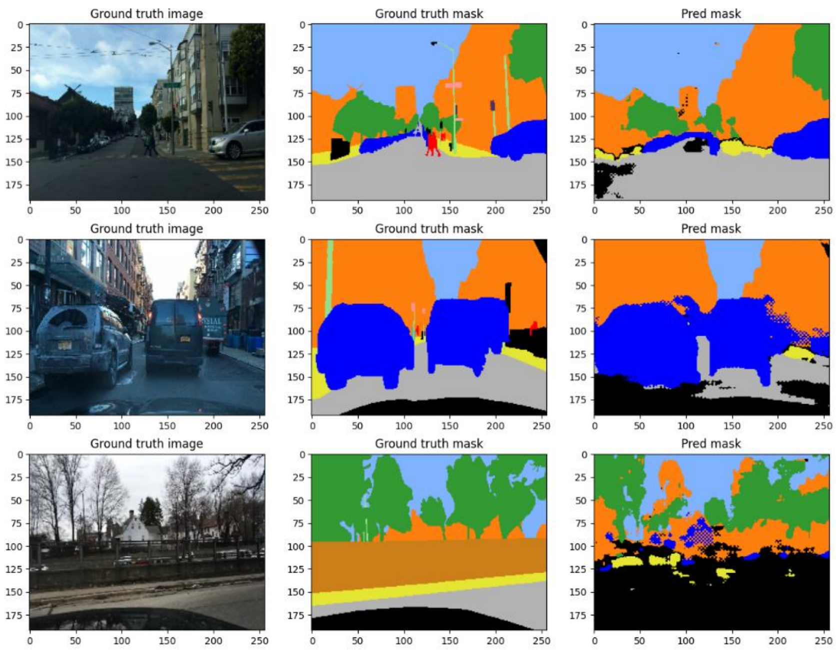


Figure 5. Qualitative results of semantic segmentation using the BDD100K tracking dataset.

Performance on the BDD100K dataset, which involves ground-level dashboard footage in diverse traffic conditions, follows a similar trend. The proposed method records the highest MOTA of 60.7, showing a consistent margin of improvement over CFTFormer (58.2), Decoupled Motion Pyramid (56.3), and SuperMOT (55.6), as detailed in Table 4. The IDF1 score rises to 61.8, suggesting strong identity preservation despite challenges such as rapid illumination shifts, occlusions near intersections, and sudden object scale variations. The HOTA score peaks at 47.9, surpassing all competing models and reflecting the system's ability to maintain long-term object associations while minimizing fragmentation. In terms of computational efficiency and tracking stability, the framework again achieves lowest false positives (FP = 4105) and minimum ID switches (IDsw = 489), reinforcing its robustness across heterogeneous urban scenarios.

**Journal of Information Systems Engineering and Management**
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
**Research Article**

Table 4 – Results on BDD100K Dataset

| Method | MOTA ↑ | IDF1 ↑ | HOTA ↑ | FP ↓ | FN ↓ | IDsw ↓ |
|---|---|---|---|---|---|---|
| DeepSORT | 48.2 | 51.7 | 40.3 | 5900 | 82150 | 8842 |
| STRAM [47] | 45.55 | 55.9 | - | - | - | **3056** |
| CFTFormer [19] | 59.8 | 67.7 | 46.3 | 4301 | 73800 | 5273 |
| SAM2MOT [48] | 58 | 71 | - | - | - | 5184 |
| Proposed | **60.7** | **68.8** | **47.9** | **4105** | **72085** | 4893 |

To understand the individual contributions of the HM-SP and BTAT modules, an ablation study was conducted under four configurations: baseline (ResNet18 only), ResNet18 with HM-SP, ResNet18 with BTAT, and the full pipeline combining all three. Results, presented in Table 5, show that the baseline configuration, which utilizes only the ResNet18 semantic backbone, yields a MOTA of 49.2 and IDF1 of 51.0 on VisDrone-MOT, with similar underperformance on BDD100K. When the HM-SP module is added, MOTA improves to 54.5, and IDF1 climbs to 56.3, demonstrating that motion-aware semantic pruning effectively suppresses background clutter and enhances object focus, particularly in occlusion-prone frames. Similarly, integrating the BTAT module alone raises MOTA to 53.6 and IDF1 to 55.0, validating the utility of bidirectional temporal modeling in reinforcing frame-to-frame identity alignment.

The most substantial gains are observed when both modules are jointly enabled. Full architecture achieves the highest metrics across all categories, with MOTA = 58.3, IDF1 = 60.2, and HOTA = 45.3 on VisDrone-MOT, and MOTA = 59.8, IDF1 = 61.8, and HOTA = 47.9 on BDD100K. The combined reduction in false positives, false negatives, and ID switches indicate that HM-SP and BTAT modules work synergistically, with HM-SP reducing noise and BTAT ensuring temporal alignment and re-identification. These results establish that the proposed pipeline not only performs better than existing state-of-the-art models but also demonstrates the benefit of modular motion pruning and temporal attention.

While this study primarily emphasizes quantitative improvements, qualitative observation of tracking results further reinforces these findings. In multiple VisDrone-MOT sequences with occluded pedestrians or motorcycles moving in and out of the frame, the proposed method consistently preserves identity labels, unlike DeepSORT and BYTETrack, which frequently experience ID-switches. Similarly, in night-time driving sequences of BDD100K, where YOLO-based detectors tend to miss small objects, the hybrid pruning strategy retains semantically strong, motion-verified features that allow BTAT to recover identities even after temporary disappearance.

## 5.3 ABLATION STUDY

Table 5 presents an ablation study evaluating the proposed modules within the overall architecture using both the VisDrone-MOT and BDD100K datasets. The baseline configuration, which includes only the ResNet18 backbone without any explicit spatial or temporal enhancements, achieves modest performance levels, with a MOTA of 49.2% and IDF1 of 51.0% on VisDrone-MOT, and 50.1% and 52.4% respectively on BDD100K. These values indicate the fundamental capability of the feature extractor to identify objects, but also reveal its limitations in handling motion consistency, object occlusion, and background noise, as evidenced by elevated false positive (FP), false negative (FN), and identity switch (IDsw) counts.

Incorporating only the HM–SP module into the pipeline leads to a clear performance enhancement across all metrics. On VisDrone-MOT, MOTA improves from 49.2% to 54.5%, while IDF1 increases to 56.3%, and HOTA improves from 38.7 to 42.5. A similar trend is observed in BDD100K, where MOTA reaches 55.1%, and IDF1 increases to 57.4%. These improvements validate the spatial pruning mechanism's ability to suppress static background regions by leveraging motion confidence and semantic cues, thereby facilitating the model's focus on dynamic and semantically relevant object features. Furthermore, the sharp reduction in identity switches and false detections confirms that HM–SP enhances spatial discrimination and continuity in object representations across frames.

When the BTAT module is applied independently, performance gains are also observed, although slightly less pronounced compared to HM–SP. On VisDrone-MOT, MOTA rises to 53.6% and IDF1 to 55.0%, while BDD100K

**Journal of Information Systems Engineering and Management**
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

sees improvements to 54.3% and 56.1%, respectively. These results indicate that bidirectional attention applied over temporal embeddings effectively captures motion dependencies in both preceding and succeeding frames, leading to better temporal alignment and object identity preservation. The BTAT module contributes to mitigating tracking failures caused by occlusions or sudden motion shifts by attending to long-range temporal cues, although the absence of spatial pruning limits its ability to handle noisy backgrounds effectively.

The integration of both modules into the full pipeline yields the highest accuracy and robustness across all evaluated dimensions. The combined configuration achieves a MOTA of 58.3%, IDF1 of 60.2%, and HOTA of 45.3 on VisDrone-MOT, while on BDD100K, it achieves 59.8%, 61.8%, and 47.9, respectively. Notably, the number of false positives and identity switches is significantly reduced, highlighting the synergistic benefits of integrating spatial and temporal attention mechanisms.

Table 5 – Ablation Study: HM-SP + BTAT

| **Configuration** | **Dataset** | **MOTA ↑** | **IDF1 ↑** | **HOTA** | **FP ↓** | **FN ↓** | **IDsw ↓** |
|---|---|---|---|---|---|---|---|
| Baseline (ResNet18 only) | VisDrone-MOT | 49.2 | 51.0 | 38.7 | 4980 | 81230 | 715 |
| Baseline (ResNet18 only) | BDD100K | 50.1 | 52.4 | 39.5 | 4810 | 79745 | 682 |
| + HM-SP only | VisDrone-MOT | 54.5 | 56.3 | 42.5 | 4200 | 76650 | 530 |
| + HM-SP only | BDD100K | 55.1 | 57.4 | 43.1 | 4025 | 75010 | 509 |
| + BTAT only | VisDrone-MOT | 53.6 | 55.0 | 41.2 | 4350 | 77450 | 545 |
| + BTAT only | BDD100K | 54.3 | 56.1 | 42.3 | 4200 | 76030 | 525 |
| HM-SP + BTAT (Proposed Full Pipeline) | VisDrone-MOT | **58.3** | **60.2** | **45.3** | **3652** | **74215** | **451** |
| HM-SP + BTAT (Proposed Full Pipeline) | BDD100K | **59.8** | **61.8** | **47.9** | **4105** | **72085** | **489** |

## 6. CONCLUSION

The development of the proposed ResNet18 + HM-SP + BTAT framework was preceded by extensive experimentation with existing tracking architectures and temporal modeling strategies. Several baseline models, although effective in certain scenarios, revealed limitations when deployed in challenging environments involving occlusions, small object detection, and dynamic background interference. These limitations guided architectural decisions and highlighted the need for a more adaptive and temporally consistent pipeline.

In the initial phase, baseline tracking models such as YOLOv8 combined with DeepSORT were evaluated due to their proven performance in object detection and real-time tracking. While YOLOv8 delivered high frame-wise detection precision, its lack of intrinsic temporal awareness led to severe degradation in ID consistency. In night-time and fast-motion VisDrone-MOT sequences, the model incurred frequent identity switches and fragmentations, especially when objects temporarily exited the frame or underwent scale shifts. Post-hoc integration with DeepSORT attempted to mitigate these issues using motion and appearance-based Kalman filtering, but the overall MOTA and IDF1 scores remained substantially lower than desired. Specifically, as seen in Table 3, the IDF1 for YOLOv8-DeepSORT configurations stagnated below 45 on VisDrone-MOT, limiting its suitability for drone-based tracking. Similar issues were noted with CFTFormer, which, despite its transformer-based design, exhibited high memory consumption and poor runtime stability due to dense feature fusion without motion-specific filtering. These observations revealed that spatial detection accuracy alone was insufficient for robust object tracking in highly dynamic environments.

**Journal of Information Systems Engineering and Management**
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
**Research Article**

To improve temporal consistency, GRU and LSTM-based recurrent networks were experimented with as replacements for traditional tracking heads. These methods showed promise in short-range dependency modeling but ultimately underperformed due to their limited memory window and sequential bottlenecks. In multi-object tracking tasks involving rapid object re-entries and partial occlusions, RNNs failed to re-establish identity continuity over long sequences. The gradient flow through time quickly deteriorated, resulting in either tracking drift or complete identity reassignment. Additionally, training these architectures introduced optimization instability when batch sizes were constrained by GPU memory. These challenges highlighted the need for parallelized and bidirectional temporal modeling, ultimately motivating the shift towards transformer-based reasoning in the proposed BTAT module.

Another persistent challenge was related to small object occlusion and semantic clutter, particularly in aerial footage from the VisDrone-MOT dataset. Baseline models processed full-frame feature maps, making no distinction between motion-relevant foreground features and static background content. As a result, low-contrast or partially occluded objects, such as pedestrians crossing under trees or vehicles at a distance, were often missed or misclassified. This led to elevated false negatives and identity loss. Attempts to increase input resolution or incorporate background subtraction only marginally improved accuracy but significantly slowed inference. The turning point was the introduction of Hybrid Motion–Semantic Pruning (HM-SP), which enabled selective suppression of static and low-motion regions by fusing motion confidence with semantic saliency. This pruning approach not only reduced computational load but also enhanced object-level focus, directly improving MOTA and IDF1 as observed in Table 5.

Eventually, after iterating through multiple backbone architectures, detection heads, and temporal models, the final convergence was achieved with the ResNet18 + HM-SP + BTAT pipeline. The decision to adopt ResNet18 was based on its balance of computational efficiency and semantic representation strength, which allowed deployment on edge devices without compromising accuracy. The HM-SP module provided lightweight motion-aware filtering that refined input features before temporal modeling. Meanwhile, the BTAT module, with its bidirectional self-attention mechanism, addressed the limitations of RNNs by modeling both past and future frames for better identity preservation and occlusion handling. Together, this configuration led to consistent performance gains across both datasets, achieving the highest scores in MOTA, IDF1, and HOTA with the lowest number of identity switches and false detections, as comprehensively demonstrated in Tables 3, 4, and 5.

## 7. CONCLUSION

This study introduced an end-to-end multi-object tracking framework that seamlessly integrates spatial, motion, and temporal cues to enhance tracking performance in complex visual environments. By integrating a lightweight ResNet18 with an HM-SP module and BTAT, the system effectively tackles persistent object tracking challenges like occlusion, background clutter, scale changes, and rapid transitions. The use of motion-guided semantic pruning allows the system to discard irrelevant background features, reducing redundancy and enhancing discriminability. The bidirectional temporal attention ensures robust identity preservation, even when objects temporarily exit the frame or change appearance drastically. This combination results in a compact, interpretable, and highly effective tracking pipeline suitable for deployment in real-time, resource-constrained environments such as UAV platforms and smart surveillance systems. The incorporation of these components within a unified modular framework resulted in reliable identity association, more rapid convergence, and enhanced robustness against visual disturbances, thereby illustrating both technical efficacy and practical applicability.

This work advances multi-object tracking by introducing dynamic feature filtering guided by motion priors, addressing a gap in recent models. Traditional tracking frameworks often process spatial features uniformly, allowing static background noise to propagate through temporal modules, thereby degrading identity coherence and increasing computational overhead. The proposed HM-SP module directly addresses this by combining motion confidence with semantic saliency to focus on relevant foreground dynamics. Simultaneously, BTAT advances temporal reasoning beyond LSTM-based models by attending to both past and future contexts without sequential bottlenecks. Together, these modules provide a novel architectural synergy that outperforms motion-agnostic transformer models and frame-wise trackers. This research thus extends the MOT literature by demonstrating that motion-aware pruning is not merely a performance enhancement, but a structural necessity for scalable, identity-consistent tracking under real-world constraints. The framework excels in high-impact real-time uses, including UAV-based surveillance in challenging environments where stable tracking is needed, smart traffic systems for consistent vehicle and pedestrian identification in busy scenes, and crowd analytics for monitoring safety in dense public spaces.

Despite its effectiveness, the current framework does have certain limitations. The model relies primarily on 2D spatial and motion information, which can lead to reduced accuracy in environments with dense fog, rain, or depth ambiguity, where semantic signals are weak or distorted. Furthermore, the temporal reasoning in BTAT, while robust, is constrained by a fixed context window, making it less effective for long-term identity preservation in very long or discontinuous video sequences. Additionally, while HM-SP significantly reduces feature map noise, it does not explicitly handle cross-camera re-identification or domain adaptation, which limits its scalability to multi-view or cross-domain deployments.

Future research will explore several directions to further enhance this framework. First, integrating depth-aware fusion or monocular depth estimation could enable more accurate occlusion handling and improve spatial reasoning under poor visibility. Second, vision-language grounding could be explored to support target-specific tracking based on semantic queries, enabling more selective surveillance. Third, adapting the model for multi-UAV collaborative tracking scenarios would extend its utility in large-scale surveillance networks. Finally, to enhance deployment viability, future work will focus on quantization, pruning, and hardware-specific optimization, targeting platforms such as NVIDIA Jetson or Intel Movidius for real-time inference under stringent power and memory budgets. These enhancements will make the proposed system even more applicable in mission-critical and embedded vision scenarios.

## DECLARATIONS

The authors declare that there are no conflicts of interest related to the publication of this paper. The research work was conducted independently and without any financial or personal relationships that could be perceived as potential sources of bias.

## REFERENCES

[1] Du, C., Lin, C., Jin, R., Chai, B., Yao, Y., & Su, S. (2024). Exploring the state-of-the-art in multi-object tracking: A comprehensive survey, evaluation, challenges, and future directions. *Multimedia tools and applications*, *83*(29), 73151-73189.

[2] Pal, S. K., Pramanik, A., Maiti, J., & Mitra, P. (2021). Deep learning in multi-object detection and tracking: state of the art. *Applied Intelligence*, *51*(9), 6400-6429.

[3] Hosain, M. T., Zaman, A., Abir, M. R., Akter, S., Mursalin, S., & Khan, S. S. (2024). Synchronizing object detection: Applications, advancements and existing challenges. *IEEE access*, *12*, 54129-54167.

[4] Bisio, I., Garibotto, C., Haleem, H., Lavagetto, F., & Sciarrone, A. (2023). Vehicular/non-vehicular multi-class multi-object tracking in drone-based aerial scenes. *IEEE Transactions on Vehicular Technology*, *73*(4), 4961-4977.

[5] Pargal, S., Das, D., Sahoo, B., Mitra, B., & Chakraborty, S. (2024). GRIDS: Personalized guideline recommendations while driving through a new city. *ACM Transactions on Recommender Systems*, *2*(2), 1-28.

[6] Guan, Z., Wang, Z., Zhang, G., Li, L., Zhang, M., Shi, Z., & Jiang, N. (2025). Multi-object tracking review: retrospective and emerging trend. *Artificial Intelligence Review*, *58*(8), 235.

[7] Yu, W., Liu, R., Chen, D., & Hu, Q. (2024). Explainability enhanced object detection transformer with feature disentanglement. *IEEE Transactions on Image Processing*.

[8] Madjid, N. A., Ahmad, A., Mebrahtu, M., Babaa, Y., Nasser, A., Malik, S., ... & Khonji, M. (2025). Trajectory prediction for autonomous driving: Progress, limitations, and future directions. *arXiv preprint arXiv:2503.03262*.

[9] Fazli, M. S., & Quinn, S. (2024). Object Tracking in a \$360^ o \$ View: A Novel Perspective on Bridging the Gap to Biomedical Advancements. *arXiv preprint arXiv:2412.01119*.

[10] Tang, X., Zhang, Z., & Qin, Y. (2021). On-road object detection and tracking based on radar and vision fusion: A review. *IEEE Intelligent Transportation Systems Magazine*, *14*(5), 103-128.

[11] Koonce, B. (2021). ResNet 50. In *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization* (pp. 63-72). Berkeley, CA: Apress.

**Journal of Information Systems Engineering and Management**
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
**Research Article**


[12] Koonce, B. (2021). EfficientNet. In *Convolutional neural networks with swift for Tensorflow: image recognition and dataset categorization* (pp. 109-123). Berkeley, CA: Apress.

[13] Zhu, Y., & Newsam, S. (2017, September). Densenet for dense flow. In *2017 IEEE international conference on image processing (ICIP)* (pp. 790-794). IEEE.

[14] Sun, G., Liu, Y., Ding, H., Wu, M., & Van Gool, L. (2024). Learning local and global temporal contexts for video semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(10), 6919-6934.

[15] Liu, H. I., Galindo, M., Xie, H., Wong, L. K., Shuai, H. H., Li, Y. H., & Cheng, W. H. (2024). Lightweight deep learning for resource-constrained environments: A survey. *ACM Computing Surveys*, *56*(10), 1-42.

[16] Wang, T., Wu, L., & Sun, C. (2022). A coarse-to-fine approach for dynamic-to-static image translation. *Pattern Recognition*, *123*, 108373.

[17] Bisio, I., Garibotto, C., Haleem, H., Lavagetto, F., & Sciarrone, A. (2023). Vehicular/non-vehicular multi-class multi-object tracking in drone-based aerial scenes. *IEEE Transactions on Vehicular Technology*, *73*(4), 4961-4977.

[18] Muhammad, K., Hussain, T., Ullah, H., Del Ser, J., Rezaei, M., Kumar, N., ... & De Albuquerque, V. H. C. (2022). Vision-based semantic segmentation in scene understanding for autonomous driving: Recent achievements, challenges, and outlooks. *IEEE Transactions on Intelligent Transportation Systems*, *23*(12), 22694-22715.

[19] Amirkhani, A., & Khoshnevis, S. A. (2025). CFTformer: End-to-End Cross-Frame Multi-Object Tracking With Transformer. *IEEE Access*.

[20] Du, C., Lin, C., Jin, R., Chai, B., Yao, Y., & Su, S. (2024). Exploring the state-of-the-art in multi-object tracking: A comprehensive survey, evaluation, challenges, and future directions. *Multimedia tools and applications*, *83*(29), 73151-73189.

[21] Han, S., Huang, P., Wang, H., Yu, E., Liu, D., & Pan, X. (2022). Mat: Motion-aware multi-object tracking. *Neurocomputing*, *476*, 75-86.

[22] Hung, W. C., Kretzschmar, H., Lin, T. Y., Chai, Y., Yu, R., Yang, M. H., & Anguelov, D. (2020). Soda: Multi-object tracking with soft data association. *arXiv preprint arXiv:2008.07725*.

[23] Cao, J., Song, C., Song, S., Xiao, F., Zhang, X., Liu, Z., & Ang Jr, M. H. (2021). Robust object tracking algorithm for autonomous vehicles in complex scenes. *Remote Sensing*, *13*(16), 3234.

[24] Zhou, Z., Li, X., Zhang, T., Wang, H., & He, Z. (2021). Object tracking via spatial-temporal memory network. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(5), 2976-2989.

[25] Zhang, H., Qu, S., Li, H., Xu, W., & Du, X. (2022). A motion-appearance-aware network for object change detection. *Knowledge-Based Systems*, *255*, 109612.

[26] Yao, J., Zhang, J., Wang, Z., & Shao, L. (2022). Semantic and context features integration for robust object tracking. *IET Image Processing*, *16*(5), 1268-1279.

[27] Liu, L., Cheng, Y., Chen, D., He, J., Liò, P., Schönlieb, C. B., & Aviles-Rivero, A. I. (2023). Traffic video object detection using motion prior. *arXiv preprint arXiv:2311.10092*.

[28] Su, J., Wu, C., & Yang, S. (2023). Object-Tracking Algorithm Combining Motion Direction and Time Series. *Applied Sciences*, *13*(8), 4835.

[29] Yang, Y., Deng, Y., Zhang, J., Gu, H., & Dong, Z. (2024). SiamMo: Siamese Motion-Centric 3D Object Tracking. *arXiv preprint arXiv:2408.01688*.

[30] Raj, A., Wang, L., & Gedeon, T. (2025, April). Tracknetv4: Enhancing fast sports object tracking with motion attention maps. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.

[31] Gao, J., Han, G., Zhu, H., & Liao, L. (2024). Multiple moving vehicles tracking algorithm with attention mechanism and motion model. *Electronics*, *13*(1), 242.

[32] Han, X., Oishi, N., Tian, Y., Ucurum, E., Young, R., Chatwin, C., & Birch, P. (2025). ETTrack: enhanced temporal motion predictor for multi-object tracking. *Applied Intelligence*, *55*(1), 33.

[33] Li, S., Ke, L., Yang, Y. H., Piccinelli, L., Segu, M., Danelljan, M., & Gool, L. V. (2024, September). Slack: Semantic, location, and appearance aware open-vocabulary tracking. In *European Conference on Computer Vision* (pp. 1-18). Cham: Springer Nature Switzerland.

**Journal of Information Systems Engineering and Management**
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

[34] Zhao, L., Wang, M., & Yue, Y. (2024, May). SCMOT: Improving 3D Multi-Object Tracking via Semantic Inference and Confidence Optimization. In *2024 36th Chinese Control and Decision Conference (CCDC)* (pp. 2524-2529). IEEE.

[35] Hu, B., Xia, M., Zhao, D., & Wu, G. (2025). MONA: Moving Object Detection from Videos Shot by Dynamic Camera. *arXiv preprint arXiv:2501.13183*.

[36] Ji, D., Gao, S., Zhu, L., Zhu, Q., Zhao, Y., Xu, P., ... & Ye, J. (2024). View-centric multi-object tracking with homographic matching in moving uav. *arXiv preprint arXiv:2403.10830*.

[37] Qian, Y., Wang, Z., Gao, Y., Zhang, W., & Wang, H. (2024). A Multi-Object tracking method in moving UAV based on IoU matching. *IEEE Access*.

[38] Yuan, Y., Wu, Y., Zhao, L., Liu, Y., & Pang, Y. (2024). End-to-end multiple object tracking in high-resolution optical sensors of drones with transformer models. *Scientific Reports*, *14*(1), 25543.

[39] Song, I., & Lee, J. (2024, October). SFTrack: A Robust Scale and Motion Adaptive Algorithm for Tracking Small and Fast Moving Objects. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 10870-10877). IEEE.

[40] Yang, L., Wang, Y., Kong, L., Bai, Y., & Tao, B. (2024, April). Object detection of visdrone based on attention mechanism and FasterNet. In *2024 5th International Conference on Computer Vision, Image and Deep Learning (CVIDL)* (pp. 257-261). IEEE