An Efficient Multi-Object Tracking Framework for Embedded Vision via Multi-Feature Fusion

Jianguo Zhang, Yun Chen, Wenxing Sun, Yuhui Chen, Junwen Yao, Hua Ye, and Duanjiao Li

***Abstract*—Despite significant progress in single-object detection, existing methods still face substantial technical challenges when deployed in complex and dynamic environments. Moreover, most models lack the capability to maintain stable identity information over time, leading to limited applicability in scenarios requiring persistent tracking and reliable spatiotemporal association. These limitations become even more pronounced on resource-constrained embedded platforms, where computational efficiency and real-time performance must be carefully balanced against detection precision and robustness. To address these challenges, our research presents an efficient multi-object tracking framework that combines SSD-based object detection with a multi-feature fusion association strategy. The proposed method leverages complementary appearance, motion, and geometric cues to construct a unified similarity matrix, enabling robust inter-frame correspondence and reducing identity switches under challenging conditions. A Kalman-based motion model and a complete track-management scheme further enhance robustness against short-term occlusions and intermittent detection failures. The system is implemented on a resource-constrained Raspberry Pi 3B platform, with dedicated optimizations to ensure real-time performance. Extensive experiments on the VOT2017 benchmark demonstrate that the proposed approach achieves a favorable balance between accuracy, stability, and computational efficiency. These results highlight the practicality of deploying the framework in real-world embedded vision applications and provide a strong foundation for future extensions incorporating learned temporal models and lightweight re-identification modules.**

***Index Terms*—Multi-object tracking, Object detection, Similarity-based association, Real-time tracking.**

I. Introduction

MULTI-OBJECT tracking (MOT) plays a central role in numerous computer vision applications, including intelligent surveillance, autonomous driving, human–computer interaction, and video-based behavior analysis. The primary objective of MOT is to simultaneously localize multiple targets in a video sequence and to maintain consistent identity assignments over time, thereby producing continuous and reliable spatiotemporal trajectories [1]. Despite the remarkable progress brought by deep learning in object detection and representation learning, MOT remains highly challenging due to target occlusion, abrupt motion patterns, appearance variations, and the presence of dense or cluttered scenes. These difficulties often lead to identity switches, fragmented trajectories, and compromised tracking stability.

Modern MOT systems predominantly follow the *tracking-by-detection* paradigm, where an external detector is applied to each frame, and inter-frame associations are established to form identity-consistent trajectories [2]. In such systems, the accuracy of object detection and the reliability of the association strategy jointly determine the overall performance. A key challenge lies in designing a robust and discriminative matching mechanism that can leverage both appearance cues and geometry consistency while handling imperfect or missing detections. As a result, recent efforts have focused on integrating deep appearance embeddings, motion models, and geometric relations into unified association frameworks [3]. While such integrated frameworks have pushed the state-of-the-art in controlled environments, their deployment in real-world, resource-constrained scenarios-such as on embedded systems for edge computing-introduces additional stringent requirements. These include the need for high computational efficiency, low memory footprint, and real-time inference speeds, all while maintaining acceptable tracking accuracy and robustness. Consequently, there is a growing need to develop MOT solutions that not only address the fundamental challenges of association under uncertainty but are also expressly designed for practical implementation under strict hardware limitations. This work is particularly motivated by the gap between increasingly complex, computationally heavy models and the practical demands of deployable systems.

Motivated by these observations, this research presents a comprehensive MOT system built upon SSD-based online object detection and a multi-feature fusion association strategy. The proposed framework first applies a pre-trained Single Shot MultiBox Detector (SSD) to extract high-confidence object candidates from each frame, which are then used to initialize identity-specific trackers [4]. For each subsequent frame, the system performs motion prediction using Kalman filtering, extracts appearance representations from the detected bounding boxes, and computes geometric consistency measures based on predicted and observed box shapes. These complementary cues are jointly formulated into a unified similarity matrix, which drives a bipartite matching process to associate detections with

Manuscript received February 1, 2026; revised March 13, 2026. Date of publication July 8, 2026. Date of current version July 8, 2026.

J. Zhang, and H. Ye are with the Shenzhen Institute of Artificial Intelligence and Robotics for Society, China (e-mails: zhangjianguo@cuhk.edu.cn, yehua@cuhk.edu.cn).

Yun Chen, W. Sun, Y. Chen, J. Yao, and D. Li are with the Guangdong Power Grid Co., Ltd, China (e-mails: 18026707368@163.com, swx_hust@163.com, cyhchenyuhui2023@163.com, 867271037@qq.com, LDJCYL@sina.com). Corresponding author: Duanjiao Li.

This research was supported by the Guangdong Power Grid Corporation (Grant No. GDKJXM20231474) and Longgang District Shenzhen's Ten Action Plan for Supporting Innovation Projects (Grants LGKCSDPT2024002, 2024003, 2024004).

Digital Object Identifier (DOI): 10.24138/jcomss-2026-0006

existing trajectories. New objects are dynamically initialized when unmatched detections exceed a confidence threshold, whereas lost tracks are retained and eventually terminated based on a missing-frame criterion.

The contributions can be summarized as follows:

- We develop an online multi-object tracking framework that leverages SSD for real-time object localization and efficient trajectory initialization.
- We propose a multi-feature fusion association mechanism that integrates appearance similarity, motion consistency, and shape compatibility into a unified similarity formulation, enabling robust inter-frame correspondence.
- We design a complete track management strategy, including prediction, update, re-identification, and termination, which enhances performance in challenging scenarios such as occlusions and intermittent detections.

Extensive experiments demonstrate that the proposed system provides stable and continuous tracking performance across diverse and complex video scenes, highlighting its effectiveness and applicability to real-world vision tasks.

II. Related Work

A. Deep Learning-Based Object Detection

Deep convolutional neural networks have substantially advanced the state of the art in object detection. Existing detectors can be broadly categorized into two-stage models, such as Faster R-CNN, and single-stage models, such as YOLO and SSD [5], [6]. Among them, the Single Shot MultiBox Detector (SSD) achieves an attractive balance between accuracy and computational efficiency by predicting object categories and bounding boxes from multi-scale feature maps in a single forward pass [7]. Owing to its real-time performance and strong generalization ability, SSD has become a widely adopted component in online multi-object tracking frameworks [8], [9]. In the proposed system, SSD serves as the backbone detector that provides high-confidence candidate regions for both initialization and frame-by-frame association [10].

B. Multi-Object Tracking and Tracking-by-Detection

Most contemporary MOT approaches follow the *tracking-by-detection* paradigm, in which objects are first detected and then associated across frames using motion or appearance cues [11]. Traditional approaches relied heavily on handcrafted features, spatial proximity, and simple motion models [12], but these methods typically struggled with occlusions, crowded scenes, and large appearance variations. Motivated by the limitations of early techniques, recent research has increasingly integrated deep appearance embeddings [13], [14], Kalman-based motion prediction, and geometric consistency into unified association strategies. Representative methods such as DeepSORT, FairMOT, and ByteTrack have demonstrated that jointly leveraging motion dynamics and discriminative visual representations can significantly reduce identity switches and improve long-term tracking robustness.

C. Feature Fusion for Data Association

A central challenge in MOT is the design of reliable criteria for associating detected objects with existing trajectories [15]. Deep appearance embeddings have proven essential for distinguishing targets with similar shapes or motion patterns [16], [17], especially in crowded scenes. Motion-based predictions, typically parameterized through Kalman filters, provide prior estimates of spatial location and facilitate association when visual cues are ambiguous or temporally missing [18], [19]. In addition, geometric attributes such as bounding box IoU, aspect ratio, and area ratios offer complementary information that helps constrain the association space and mitigate errors caused by partial occlusions [20]. Many recent works have demonstrated that multi-feature fusion, combining appearance, motion, and geometry cues, leads to more reliable association decisions [21]. Following this line of research, the proposed method constructs a unified similarity matrix that synthesizes these heterogeneous cues, enabling robust bipartite matching and improving performance in challenging tracking scenarios and practical applications [22].

D. Track Management in Online MOT Systems

Track initialization, update, and termination are crucial for stable long-term tracking [23], [24]. State estimation is commonly performed using Kalman filters, which model object dynamics and propagate motion states forward in time [25]. To handle short-term occlusions or intermittent detection failures, online MOT systems often adopt a track survival mechanism, where missing trajectories are temporarily maintained and reactivated if matched in future frames [26], [27]. Tracks are typically terminated once their missing-frame count exceeds a predefined threshold [28]. This paradigm has been extensively validated in numerous online MOT systems, including DeepSORT and related extensions. In line with these practices, the proposed method incorporates a complete track life-cycle management strategy that improves robustness under occlusion, clutter, and sporadic detector failures [29].

In summary, the proposed approach builds upon recent advances in deep object detection, multi-feature fusion for data association, and robust online track management. By integrating SSD detection with appearance–motion–shape similarity estimation and a principled tracking framework, our system achieves reliable and identity-consistent tracking performance across diverse and complex video scenes.

III. Method

This section presents the proposed multi-object tracking (MOT) framework from both mathematical and algorithmic perspectives (as shown in Fig. 1). Beyond formal symbol definitions, we provide refined descriptions regarding the operational mechanisms of each module and their respective contributions to tracking reliability, accuracy, and robustness. The goal is to present a logically coherent and linguistically diversified exposition that adheres to academic standards.

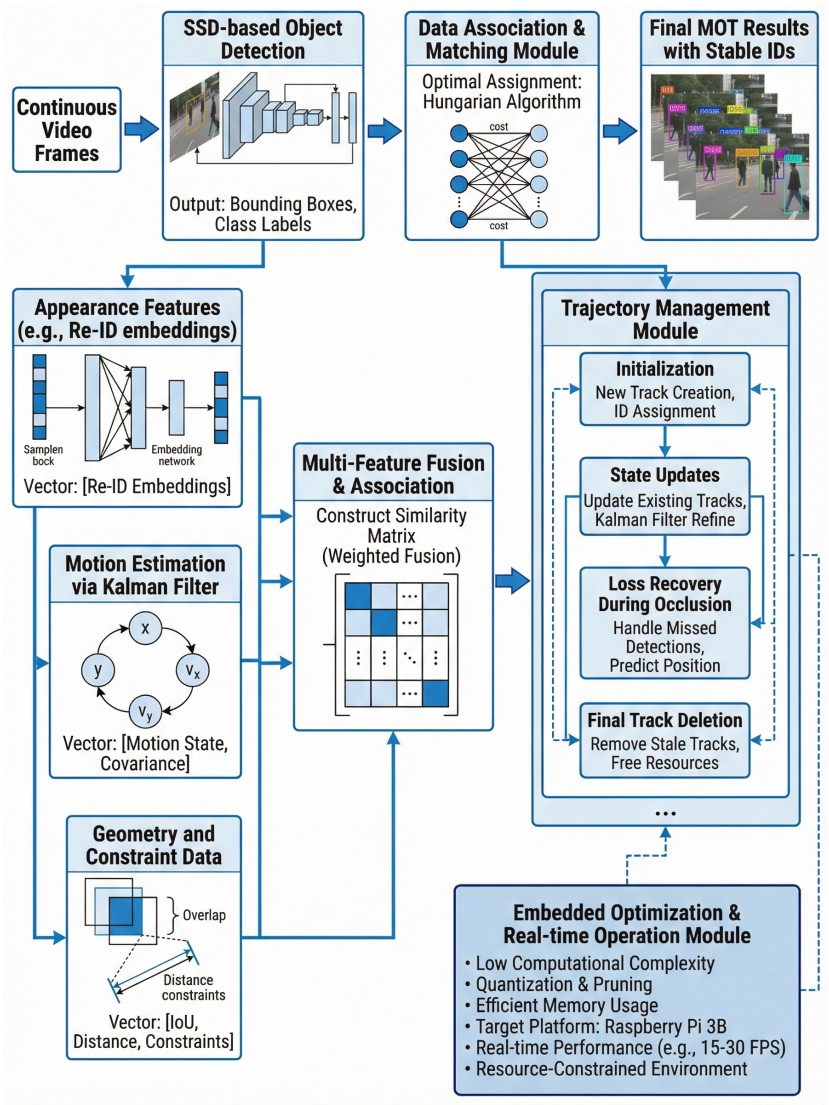

Fig. 1. The system integrates SSD-based object detection with multi-object tracking to enable real-time intelligent alarming. Detected targets are assigned unique IDs and tracked across frames based on IoU and area comparison, and an alarm is triggered when predefined overlap and time thresholds are satisfied. This iterative process ensures continuous tracking and improves the accuracy and responsiveness of the surveillance system.

A. Notation and Problem Definition

We assume that a video sequence consisting of N consecutive frames captured over time can be denoted by:

$$V = I_{t\,t=1}^{N}, \tag{1}$$

where I_t represents the video frame at time index t, and $t \in 1, 2, \ldots, N$. The task of MOT is to simultaneously localize and consistently identify multiple dynamic objects over time, ensuring both spatial accuracy and identity continuity despite environmental disturbances such as occlusion, illumination change, and object interaction. The trajectory of the i-th object is defined as the ordered sequence:

$$\tau^i = (t, ID^i, b_t^i)_{t=t_s^i}^{t_e^i}, \tag{2}$$

where ID^i corresponds to a unique identifier associated with object i, while b_t^i denotes its bounding box at time t. The indices t_s^i and t_e^i indicate the temporal span during which the object is actively tracked. This formulation enables precise temporal modeling of object dynamics throughout the sequence. The bounding box is parameterized by:

$$b_t^i = (x_t^i, y_t^i, w_t^i, h_t^i), \tag{3}$$

with (x_t^i, y_t^i) specifying the geometric center, and w_t^i, h_t^i denoting width and height. Such representation supports efficient computation of spatial relations and facilitates seamless integration with motion models. At each time step t, the system maintains a collection of active trackers:

$$T_t = \{T_t^i | i = 1, 2, 3, ..., K_t\}, \tag{4}$$

where K_t corresponds to the number of active targets. Each tracker encapsulates comprehensive descriptive information, enabling the tracker to function as a unified representation that integrates spatial dynamics and appearance characteristics. Specifically, each tracker T_t^i contains:

- A persistent identity label ID^i ensuring temporal coherence across frames.
- A motion state vector $x*t^i \in \mathbb{R}^n$ and covariance matrix $\Sigma_t^i \in \mathbb{R}^{n\times n}$, jointly modeling target dynamics via a Kalman filter.
- An appearance embedding $a*t^i \in \mathbb{R}^d$ capturing high-level visual semantics.
- A temporal history of bounding boxes $b*\tau^i*\tau = t_s^{i\,t}$ enabling consistency verification.
- A missing-frame counter m_t^i controlling lifecycle management of each track.

In conclusion, the designed structured organization allows for synergistic exploitation of multi-modal information, thereby reinforcing resilience under complex scenarios with dense targets and abrupt motion changes.

B. Video Input and Initial Detection

For the initial frame I_1, a pre-trained SSD, denoted as $\mathcal{F}*$ SSD, is employed to produce candidate object hypotheses:

$$\mathcal{D}*1 = \mathcal{F}*\text{SSD}(I_1) = (b_1^k, c_1^k, s_1^k)*k = 1^{M_1}, \quad (5)$$

where b_1^k represents the k-th bounding box, c_1^k its corresponding category label, and $s_1^k \in [0,1]$ the associated confidence score. Detections not satisfying $s_1^k \geq \theta_{\text{det}}$ are eliminated, thereby reducing the influence of spurious responses and stabilizing subsequent tracking stages. The confidence threshold θ_{det} is selected empirically on a validation subset to balance detection recall and false positives. If θ_{det} is set too low, noisy detections may trigger unstable tracker initialization and increase association ambiguity; if it is set too high, true objects may be missed or initialized too late. In our implementation, θ_{det} is chosen to favor stable initialization while preserving sufficient sensitivity to newly appearing targets. This detection phase establishes the spatial groundwork for trajectory initialization. Its quality directly influences the reliability of identity assignment and subsequently impacts long-term tracking fidelity. Each retained detection (b_1^k, s_1^k) leads to the creation of a new tracker, initialized as follows:

$$ID^i = i, \quad x_1^i = g(b_1^k), \quad \Sigma_1^i = \Sigma_0, \quad (6)$$

$$a_1^i = \phi(I_1, b_1^k), \quad m_1^i = 0, \quad (7)$$

where $g(\cdot)$ maps spatial coordinates into the Kalman state space, Σ_0 denotes the initial covariance, and $\phi(\cdot)$ extracts discriminative appearance features. This initialization procedure equips each tracker with a balanced prior, supporting rapid adaptation and accurate prediction from early frames. For each frame $t > 1$, detection is performed again to refresh spatial hypotheses:

$$\mathcal{D}*t = \mathcal{F}*\text{SSD}(I_t) = (b_t^k, c_t^k, s_t^k)_{k=1}^{M_t}. \quad (8)$$

Continuous re-detection facilitates effective recovery from transient tracking failure, compensates for inaccuracies in motion prediction, and mitigates cumulative drift, thereby increasing overall temporal robustness.

C. Multi-Feature Fusion for Data Association

To reliably associate detections with existing trackers, the proposed method integrates motion, appearance, and geometric cues through a unified fusion mechanism. This multi-dimensional strategy substantially enhances discriminative power while reducing ambiguity in densely populated scenes. Motion evolution is estimated via Kalman prediction:

$$\hat{x}*t^i = Ax*t-1^i, \quad \hat{\Sigma}*t^i = A\Sigma*t-1^i A^\top + Q, \quad (9)$$

where A is the state transition matrix and Q the process noise covariance. This predictive constraint provides temporal continuity, enabling sustained tracking even during short-term occlusion or detection dropouts. Appearance descriptors are extracted as:

$$a*t^k = \phi(I_t, b_t^k), \quad (10)$$

with similarity measured using cosine distance:

$$S_a(i,k) = \frac{a*t-1^i \cdot a*t^k}{|a*t-1^i|_2 |a_t^k|_2}. \quad (11)$$

This appearance constraint reinforces identity discrimination, particularly effective when spatial overlap among objects becomes significant. Spatial alignment is evaluated using Intersection-over-Union:

$$S_m(i,k) = \text{IoU}(\hat{b}_t^i, b_t^k). \quad (12)$$

This measure suppresses implausible associations and enforces geometric consistency in temporal correspondence. Define the width-height ratio and area ratio as:

$$r_{wh}(i,k) = \frac{w(\hat{b}_t^i)/h(\hat{b}_t^i)}{w(b_t^k)/h(b_t^k)}, \quad (13)$$

$$r_A(i,k) = \frac{A(\hat{b}*t^i)}{A(b_t^k)}. \quad (14)$$

The shape similarity is formulated as:

$$S_s(i,k) = \exp(-\alpha_1 |r_{wh}(i,k)-1| - \alpha_2 |r_A(i,k)-1|). \quad (15)$$

By constraining dimensional coherence, this component mitigates mismatches originated from scale distortions or irregular detections. The overall similarity score is derived through weighted aggregation:

$$S(i,k) = \lambda_a S_a(i,k) + \lambda_m S_m(i,k) + \lambda_s S_s(i,k), \quad (16)$$

subject to $\lambda_a + \lambda_m + \lambda_s = 1$. This integrative formulation harmonizes complementary cues, yielding a more stable and reliable association decision. The cost matrix is defined as:

$$C = 1 - S. \quad (17)$$

The Hungarian algorithm determines the optimal matching set $\mathcal{M}_t$. For each matched pair (i,k), the tracker state is updated as:

$$(x_t^i, \Sigma_t^i) = \text{KalmanUpdate}(\hat{x}_t^i, \hat{\Sigma}_t^i, b_t^k), \quad (18)$$

while the appearance vector is refined via:

$$a*t^i = \beta a*t-1^i + (1-\beta)a_t^k. \quad (19)$$

This progressive update mechanism enables continuous refinement of object representation, enhancing adaptability to gradual visual changes.

For unmatched trackers, the missing counter is incremented as $m_t^i = m_{t-1}^i + 1$. If $m_t^i > \tau_{\max}$, the tracker is safely removed, preventing long-term drift. The final trajectory set is given by:

$$\mathcal{S} = \tau^{i\,K}_{i=1}, \tag{20}$$

where K denotes the total number of valid trajectories. This output preserves coherent temporal structure and serves as a reliable basis for subsequent high-level analysis tasks such as activity recognition and behavior modeling.

IV. Experiments

This section presents a series of experiments designed to evaluate the effectiveness and practical applicability of the proposed multi-object tracking framework. The experimental setup, datasets, and evaluation metrics are detailed below.

A. Experimental Settings

1) Dataset: We conduct experiments on the VOT2017 [30] benchmark, a widely recognized dataset for short-term single-object tracking evaluation. Compared with its predecessor VOT2016, the VOT2017 dataset introduces several significant enhancements. First, the composition of video sequences has been optimized by replacing ten relatively simple sequences with ten newly collected challenging ones, ensuring a more balanced distribution of attributes such as occlusion, illumination changes, motion complexity, and camera dynamics. Second, annotation quality has been substantially improved: all sequences were reannotated using pixel-accurate segmentation masks, and axis-aligned bounding boxes were re-fitted from these masks, resulting in more reliable ground-truth.

A notable feature of VOT2017 [30] is the adoption of a concealed test set consisting of 60 carefully curated video sequences. Although not publicly released, its attribute distribution reflects that of the public set, ensuring fairness and robustness in benchmarking. The evaluation protocol of VOT2017 [30] is primarily based on three core indicators-Accuracy, Robustness, and Expected Average Overlap (EAO). Among these, EAO serves as a comprehensive performance measure balancing tracking precision and stability, and is recognized as one of the most authoritative metrics in the field.

2) Evaluation Metrics: To quantitatively assess the tracking performance of the proposed method, we employ two geometry-based metrics that are commonly used in visual tracking: *Overlap Precision* and *Center Precision*. Both metrics are computed with respect to the ground-truth bounding boxes provided in the dataset.

Overlap Precision. Overlap Precision (OP) measures the spatial accuracy of the predicted bounding boxes. For each frame, the Intersection-over-Union (IoU) between the tracked bounding box and the ground-truth annotation is computed as:

$$\text{IoU} = \frac{|B_{\text{pred}} \cap B_{\text{gt}}|}{|B_{\text{pred}} \cup B_{\text{gt}}|}. \tag{21}$$

We then vary the IoU threshold from 0 to 1 and compute the percentage of frames in which the IoU exceeds the given threshold. The resulting curve provides a comprehensive evaluation of localization precision across different accuracy requirements.

Center Precision. Center Precision (CP) evaluates tracking stability by measuring the Euclidean distance between the centers of predicted and ground-truth bounding boxes: $d = \|c_{\text{pred}} - c_{\text{gt}}\|_2$. A distance threshold ranging from 0 to 50 pixels is applied, and the proportion of frames with d below each threshold is recorded. This metric is particularly sensitive to jittery motion in the tracker and therefore provides an informative assessment of temporal smoothness and stability.

Together, these two metrics offer a complementary characterization of tracking performance, capturing both bounding-box alignment accuracy and center-level temporal consistency.

3) Implementation Details: The proposed tracking framework is implemented using the KCFTrackerSDK, a lightweight software development kit specifically designed for embedded platforms. The KCFTrackerSDK integrates a Kernelized Correlation Filter (KCF) [31] backend to support robust visual tracking under constrained computational environments. All experiments are conducted on a Raspberry Pi 3B, which features a quad-core ARM Cortex-A53 CPU clocked at 700 MHz and equipped with 1 GB of RAM. Such limited hardware resources impose strict requirements on algorithmic efficiency and memory usage. The system relies on OpenCV 2.4.9 for fundamental image processing operations, matrix computations, and video stream handling. Additionally, the MTCNN library [32] is incorporated to perform object detection during the initialization stage. Since both the SSD-based initialization and feature extraction are computationally expensive, particular attention is given to optimizing memory bandwidth and floating-point computation efficiency. In particular, the KCF [31] component involves dense frequency-domain transformations and element-wise matrix operations, which require careful pipeline optimization and the utilization of ARM-specific instruction sets to maintain real-time performance on the embedded platform.

Training Strategy. The SSD detector employed in our system is initialized using a pre-trained model and subsequently fine-tuned on a representative subset of target-domain data to enhance detection accuracy and robustness under deployment conditions. Fine-tuning is conducted using stochastic gradient descent with momentum, and the learning rate is adjusted following a step-wise decay schedule to facilitate stable convergence. To improve generalization capability, a range of data augmentation strategies is applied during training, including random cropping, color jittering, and scale perturbation, which collectively increase the model's resilience to variations in object appearance and environmental conditions. Since the tracking component operates in a non-parametric manner, no additional training is required for the KCF [31] tracker itself. Since the tracking component operates in a non-parametric manner, no additional training is required for the KCF tracker itself. The appearance feature extractor is fine-tuned using detector-aligned object crops, so that the extracted representations remain consistent with the spatial regions produced by

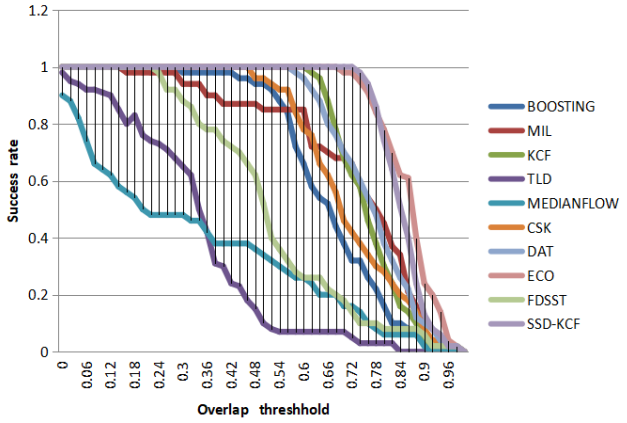

Fig. 2. Track algorithm area overlap rate statistics.

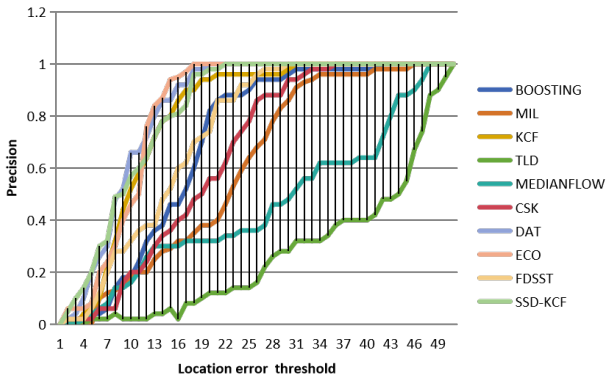

Fig. 3. Track algorithm center rate statistics.

the SSD detector during inference. This design improves the reliability of appearance-based association while avoiding the overhead of a fully end-to-end joint training framework.

Inference Procedure. During inference, each incoming frame first undergoes preprocessing, including resizing, color normalization, and optional noise suppression. The SSD detector is invoked only at initialization or re-detection events, thus reducing per-frame computational cost. For each active tracker, a Kalman prediction step is performed prior to feature extraction and similarity computation. The appearance embeddings are obtained via forward propagation through the lightweight feature extractor, while motion predictions are updated through the KCF/SSD hybrid mechanism. Finally, data association is carried out using the unified similarity matrix, and track states are updated accordingly.

To further improve deployment efficiency on Raspberry Pi 3B, we implemented four platform-oriented optimizations. (i) Batched similarity computation: instead of computing appearance, motion, and shape similarity independently for each tracker–detection pair, we organize candidate pairs into contiguous memory blocks and process them in batches. This reduces repeated memory access overhead and improves cache locality during matrix operations. (ii) Model quantization: the appearance feature extractor is quantized to reduce both memory footprint and floating-point computation cost. This is particularly beneficial on ARM-based processors with limited arithmetic throughput. (iii) Confidence-driven re-detection: the SSD detector is not invoked on every frame. Instead, re-detection is triggered when the confidence of active tracks decreases, when unmatched tracks accumulate, or when new objects are likely to enter the scene. This significantly reduces average detector overhead while preserving recovery capability. (iv) Asynchronous pipeline design: when possible, detection-related preprocessing and tracker-state updates are executed in a pipelined manner to reduce idle waiting time between stages. These optimizations are introduced specifically for embedded deployment and are not intended to maximize accuracy alone. Rather, they are designed to preserve a practical balance between responsiveness and tracking reliability under strict hardware constraints.

B. Results and Comparisons

Based on the proposed method and experimental datasets, we evaluated the performance of our approach for object tracking across various scenarios and conducted comparative analyses against several mainstream methods. The tracking algorithms were assessed primarily along three metrics:

- the overlap rate between the tracked bounding box and the ground truth annotation (IoU),
- the center location error between the tracked bounding box and the ground truth, and
- the frame rate of the tracking algorithm.

The detailed procedures and results are presented as follows.

Overlap Rate. As shown in Fig. 2, our SSD-KCF method demonstrates outstanding performance in terms of region overlap rate. Specifically, at lower overlap thresholds (0.0–0.2), the success rate of SSD-KCF approaches 1.0, performing on par with methods such as BOOSTING [33] and MIL [34]. As the overlap threshold gradually increases, the success rate of SSD-KCF declines relatively slowly. For instance, when the threshold rises to 0.6–0.7, the success rate of SSD-KCF remains significantly higher than that of several other methods, including TLD [35] and MEDIANFLOW [36]. Even at high thresholds of 0.8–0.9, where the success rate of SSD-KCF naturally decreases, it still maintains a comparatively superior level against methods like CSK [37] and DAT [38]. Overall, the success plot clearly indicates that our SSD-KCF method achieves robust performance across all overlap thresholds, exhibiting particular advantages at moderate to high levels of overlap compared to the alternative approaches.

Center Location Error. In the center location error experiments (As shown in Fig. 3), our SSD-KCF method also exhibits commendable performance. The figure shows that at low location error thresholds (1–10 pixels), the precision of SSD-KCF increases rapidly, matching the performance of leading methods such as ECO [39] and KCF [31]. As the location error threshold increases, the precision of SSD-KCF decreases gently. Within the threshold range of 10–30 pixels, the precision of SSD-KCF surpasses that of methods like TLD [35] and MEDIANFLOW [36]. When the threshold reaches 30–50 pixels, although the precision of SSD-KCF decreases, it remains at a high level and demonstrates a clear

TABLE I
THIS TEST CALCULATED THE AVERAGE FRAME RATE PER 100 CONSECUTIVE TRACKING INSTANCES.

TRACKER	FPS
BOOSTING [33]	6.1574
MIL [34]	5.15488
KCF [31]	34.3859
TLD [35]	7.32897
MEDIANFLOW [36]	57.1452
CSK [37]	33.0879
DAT [38]	9.74493
ECO [39]	1.03073
FDSST [40]	26.2778
SSD-KCF	31.4286

advantage over methods such as DAT [38] and CSK [37]. In summary, the precision plot convincingly illustrates that our SSD-KCF method delivers excellent precision control across varying location error thresholds, demonstrating strong competitiveness, especially at small to medium error ranges.

Frame Rate of Tracking. Among the tested algorithms in Table I, the MEDIANFLOW [36] algorithm leads with a frame rate of 57.15 FPS, indicating its high processing speed and ability to perform tracking tasks rapidly per frame. The CF [31] algorithm achieves 34.39 FPS, placing it in the upper-middle tier. Similarly, the CSK [37] algorithm attains 33.09 FPS, closely following KCF [31]. Our SSD-KCF algorithm runs at 31.43 FPS. While slightly lower than KCF [31] and CSK [37], it still demonstrates high efficiency sufficient for many application scenarios with real-time requirements. In contrast, the BOOSTING [33] (6.16 FPS), MIL [34] (5.15 FPS), TLD [35] (7.33 FPS), and DAT [38] (9.74 FPS) algorithms exhibit relatively lower frame rates, which may pose performance bottlenecks when processing large-scale video data or handling tasks demanding high real-time capability. Notably, the ECO [39] algorithm lags significantly behind with a mere 1.03 FPS.

Overall, regarding the frame rate metric, our SSD-KCF algorithm exhibits favourable efficiency in the object tracking experiments. It manages to process video frames at a reasonably high rate while maintaining promising tracking accuracy (as corroborated by the previous overlap and precision results), showcasing well-balanced overall performance and applicability for diverse practical scenarios. In the multiple tracking tests conducted across different datasets, the visualization of our results is shown in Fig. 4 and Fig. 5. As illustrated, the model consistently maintains high accuracy and strong stability throughout all evaluations, demonstrating the algorithm's robustness and reliability even in complex scenarios.

C. Complexity Analysis

Let K_t denote the number of active tracks and M_t the number of detections in frame t. The main computational cost of the proposed method consists of the following parts.

- *Detection:* If the SSD detector is invoked, its cost is denoted by C_{det}, which dominates the per-frame computation. Since re-detection is triggered only when necessary, the average detector cost is reduced in practice.
- *Kalman prediction and update:* For each active track, the motion prediction and correction incur linear cost with respect to the number of tracks, $O(K_t)$, assuming a fixed-dimensional state vector.
- *Similarity matrix construction:* Appearance, motion, and shape similarities are computed for each track–detection pair, resulting in $O(K_t M_t)$ pairwise operations. If the appearance embedding dimension is d, the appearance term contributes approximately $O(K_t M_t d)$.
- *Bipartite matching:* The Hungarian algorithm is applied to the cost matrix of size $K_t \times M_t$, with worst-case complexity $O(n^3)$, where $n = \max(K_t, M_t)$.
- *KCF-based local tracking:* For each maintained track, KCF performs correlation-filter-based update and localization. With FFT-based implementation, the practical cost remains low for moderate target sizes, making it suitable for embedded platforms.

Overall, the proposed framework preserves manageable computational complexity for sparse and moderately populated scenes. Compared with more computation-heavy end-to-end MOT frameworks that invoke deep detection and re-identification on every frame, the present method shifts the design toward lower average runtime and improved embedded deployability, at the cost of reduced modeling capacity for highly dynamic scenes.

Fig. 4. MOT Multi-Target Pedestrian Test Set [41].

V. DISCUSSION

The experimental results and system analyses presented provide valuable insights into the behavior, strengths, and limitations of the proposed multi-object tracking framework. This

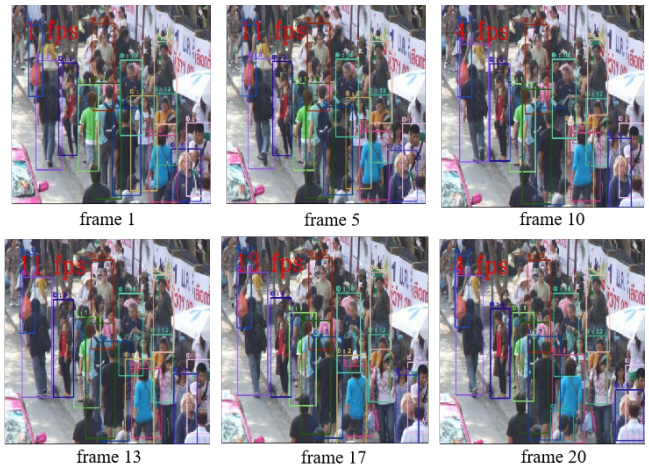

Fig. 5. CAS Large-Scale Human Scenario Testing.

section discusses these observations in detail and highlights the broader implications for real-world deployment.

A. Strengths and Performance Characteristics

The proposed method demonstrates strong performance in scenarios characterized by moderate object motion and partial occlusions. The combination of SSD-based detection and multi-feature fusion association effectively reduces identity switches and enhances temporal consistency. In particular, the appearance-motion-shape similarity integration provides a robust matching mechanism, enabling the system to maintain stable trajectories even under complex motion patterns. Moreover, the lightweight implementation of the tracking framework, coupled with platform-specific optimizations, shows that accurate visual tracking can be achieved on resource-constrained embedded devices such as the Raspberry Pi 3B. These results underline the practicality of the system for edge computing applications with limited computational resources.

B. Impact of Hardware Constraints

Despite its effectiveness, the embedded deployment environment imposes several constraints that shape the system's behavior. The limited floating-point throughput and memory bandwidth of ARM-based processors introduce bottlenecks in operations involving dense matrix computation and frequency-domain transformations (within the KCF module). Although NEON acceleration and pipeline optimization help mitigate these limitations, real-time performance remains challenging in scenes with rapidly changing dynamics or high densities of tracked objects. Furthermore, reducing the invocation frequency of the SSD detector improves efficiency but may introduce delays in detecting newly appearing objects, especially when abrupt scene changes occur.

C. Robustness and Failure Cases

The proposed system demonstrates stable performance across most evaluation scenarios. However, several failure cases were identified during experimental validation. A primary challenge arises under conditions of severe or prolonged occlusion. Although the Kalman-based motion prediction mechanism effectively bridges short-term observation gaps, extended periods without visual confirmation frequently result in track fragmentation or premature termination. Additionally, substantial variations in object appearance, particularly those induced by abrupt illumination changes or viewpoint shifts, can undermine the reliability of appearance-based similarity measures. This degradation may lead to incorrect data associations and increased identity switches in certain sequences.

Furthermore, in densely populated or highly cluttered scenes, ambiguity caused by significant bounding-box overlap can compromise the accuracy of IoU-based motion similarity estimation. Such overlap reduces the discriminative power of the geometric cue and weakens the overall effectiveness of the multi-feature fusion strategy, occasionally leading to association errors and reduced tracking consistency.

D. Limitations and Future Work

Although the proposed framework demonstrates competitive performance, several limitations suggest directions for future work. The reliance on relatively simple motion models, such as KCF and a linear Kalman filter, may limit the ability to capture complex or non-linear dynamics in highly dynamic scenarios, motivating the exploration of more expressive motion representations. In addition, the use of a fixed SSD-based detector and predefined feature extraction, while suitable for embedded deployment, may constrain further improvements in detection accuracy and identity consistency, particularly under occlusions or viewpoint changes. Overall, the system offers a reasonable trade-off between performance and computational efficiency, and the reported results provide a baseline for continued investigation into improved robustness and long-term tracking stability in dynamic or crowded environments.

The proposed framework is primarily intended for embedded vision applications that require stable online tracking under limited computational resources, rather than for highly dynamic large-scale tracking benchmarks. In particular, it is well suited to scenarios with moderate target motion, short-term occlusion, and relatively sparse to medium object density, where the combination of SSD-based detection, KCF-assisted tracking, and multi-feature association can provide a favorable balance between accuracy and speed. Typical examples include lightweight surveillance, pedestrian flow monitoring, access control, and edge-side event monitoring and alerting.

However, the simplified linear motion model may be insufficient for targets exhibiting abrupt acceleration, non-linear motion, or long-term disappearance. Likewise, in highly crowded scenes with persistent overlap, the discriminative power of motion and geometric cues may degrade, increasing identity switches. These limitations are inherent to the design trade-off in this work and are accepted to maintain real-time performance on Raspberry Pi-class hardware.

VI. Conclusion

This research presents an efficient multi-object tracking framework that combines SSD-based object detection with

a unified multi-feature association strategy integrating appearance, motion, and geometric cues. Kalman-based motion prediction and complete track lifecycle management improve robustness against short-term occlusions, missed detections, and abrupt target motion. Experiments on the VOT2017 benchmark demonstrate a favorable trade-off between tracking accuracy, robustness, and computational efficiency, and further confirm the feasibility of deployment on resource-constrained embedded platforms such as the Raspberry Pi 3B through platform-aware optimization. While performance degrades under long-term occlusion, severe appearance variation, and highly crowded scenes, future research will address these challenges via stronger temporal modeling, lightweight feature extractors, and online re-identification, as well as hardware acceleration and multi-threaded execution. Overall, the proposed framework offers a practical and extensible solution for real-world multi-object tracking applications.

REFERENCES

[1] S. Javed, M. Danelljan, F. S. Khan, M. H. Khan, M. Felsberg, and J. Matas, "Visual object tracking with discriminative filters and siamese networks: a survey and outlook," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 5, pp. 6552–6574, 2022.

[2] D. M. Jiménez-Bravo, Á. L. Murciego, A. S. Mendes, H. S. San Blás, and J. Bajo, "Multi-object tracking in traffic environments: A systematic literature review," *Neurocomputing*, vol. 494, pp. 43–55, 2022.

[3] A. Kumar, R. Vohra, R. Jain, M. Li, C. Gan, and D. K. Jain, "Correlation filter based single object tracking: A review," *Information Fusion*, vol. 112, p. 102562, 2024.

[4] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, and T.-K. Kim, "Multiple object tracking: A literature review," *Artificial intelligence*, vol. 293, p. 103448, 2021.

[5] M. Sohan, T. Sai Ram, and C. V. Rami Reddy, "A review on yolov8 and its advancements," in *International Conference on Data Intelligence and Cognitive Informatics*. Springer, 2024, pp. 529–545.

[6] Z. Cui, Y. Dai, Y. Duan, and X. Tao, "Joint object detection and multi-object tracking based on hypergraph matching." *Applied Sciences (2076-3417)*, vol. 14, no. 23, 2024.

[7] Y. Wang, K. Kitani, and X. Weng, "Joint object detection and multi-object tracking with graph neural networks," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 13 708–13 715.

[8] L. Zheng, M. Tang, Y. Chen, G. Zhu, J. Wang, and H. Lu, "Improving multiple object tracking with single object tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2453–2462.

[9] J. Cai, M. Xu, W. Li, Y. Xiong, and W. Xia, "Memot: Multi-object tracking with memory," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8090–8100.

[10] S. Li, H. Ren, X. Xie, and Y. Cao, "A review of multi-object tracking in recent times," *IET Computer Vision*, vol. 19, no. 1, p. e70010, 2025.

[11] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "Fairmot: On the fairness of detection and re-identification in multiple object tracking," *International journal of computer vision*, vol. 129, no. 11, pp. 3069–3087, 2021.

[12] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "Bytetrack: Multi-object tracking by associating every detection box," in *European conference on computer vision*. Springer, 2022, pp. 1–21.

[13] Y. Zhang, T. Wang, and X. Zhang, "Motrv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 056–22 065.

[14] F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei, "Motr: End-to-end multiple-object tracking with transformer," in *European conference on computer vision*. Springer, 2022, pp. 659–675.

[15] K. Date, G. A. Gross, and R. Nagi, "Test and evaluation of data association algorithms in hard+ soft data fusion," in *17th International Conference on Information Fusion (FUSION)*. IEEE, 2014, pp. 1–8.

[16] Y. Wu, Q. Liu, and H. Sun, "Hrtracker: multi-object tracking in satellite video enhanced by high-resolution feature fusion and an adaptive data association," *Remote Sensing*, vol. 16, no. 17, p. 3347, 2024.

[17] W. Wang, J. Li, J. Jiang, B. Wang, Q. Wang, E. Gao, and T. Yue, "Autonomous data association and intelligent information discovery based on multimodal fusion technology," *Symmetry*, vol. 16, no. 1, p. 81, 2024.

[18] Y.-L. Li, "Unsupervised embedding and association network for multi-object tracking." in *IJCAI*, 2022, pp. 1123–1129.

[19] M. P. Muresan, S. Nedevschi, and R. Danescu, "Robust data association using fusion of data-driven and engineered features for real-time pedestrian tracking in thermal images," *Sensors*, vol. 21, no. 23, p. 8005, 2021.

[20] H. Kang, L. Hou, Y. Gu, X. Lu, J. Li, and Q. Li, "Drug–disease association prediction with literature based multi-feature fusion," *Frontiers in Pharmacology*, vol. 14, p. 1205144, 2023.

[21] B. Duraisamy, T. Schwarz, and C. Wöhler, "On track-to-track data association for automotive sensor fusion," in *2015 18th International Conference on Information Fusion (Fusion)*. IEEE, 2015, pp. 1213–1222.

[22] H. Laghmara, T. Laurain, C. Cudel, and J.-P. Lauffenburger, "Heterogeneous sensor data fusion for multiple object association using belief functions," *Information Fusion*, vol. 57, pp. 44–58, 2020.

[23] Y. Xiang, A. Alahi, and S. Savarese, "Learning to track: Online multi-object tracking by decision making," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4705–4713.

[24] S. Bilakeri and K. A. Kotegar, "A review: Recent advancements in online private mode multi-object tracking," *IEEE Transactions on Artificial Intelligence*, 2025.

[25] Y.-M. Song, K. Yoon, Y.-C. Yoon, K. C. Yow, and M. Jeon, "Online multi-object tracking with gmphd filter and occlusion group management," *IEEE access*, vol. 7, pp. 165 103–165 121, 2019.

[26] Z. Guan, Z. Wang, G. Zhang, L. Li, M. Zhang, Z. Shi, and N. Jiang, "Multi-object tracking review: retrospective and emerging trend," *Artificial Intelligence Review*, vol. 58, no. 8, p. 235, 2025.

[27] D. Kang, K. Lee, C.-H. Hong, Y. Lee, J. Lee, and H. Baek, "Motas: Real-time scheduling framework for multi-object tracking capturing accuracy and stability," in *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, 2024, pp. 159–168.

[28] J. Alikhanov and H. Kim, "Online action detection in surveillance scenarios: A comprehensive review and comparative study of state-of-the-art multi-object tracking methods," *IEEE Access*, vol. 11, pp. 68 079–68 092, 2023.

[29] D. Stadler and J. Beyerer, "Improving multiple pedestrian tracking by track management and occlusion handling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 958–10 967.

[30] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. C. Zajc, T. Vojir, G. Bhat, A. Lukezic, A. Eldesokey *et al.*, "The visual object tracking vot2017 challenge results," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 1949–1972.

[31] J. F. Henriques and R. Caseiro, "High-speed tracking with kernelized correlation filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2015.

[32] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[33] H. Grabner and H. Bischof, "Real-time tracking via on-line boosting," in *Proc. British Machine Vision Conference*, 2006, pp. 47–56.

[34] B. Babenko, M.-H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," in *2009 IEEE Conference on computer vision and Pattern Recognition*. IEEE, 2009, pp. 983–990.

[35] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1409–1422, 2012.

[36] Z. Kalal, K. Mikolajczyk, and J. Matas, "Forward-backward error: Automatic detection of tracking failures," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 2756–2759.

[37] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *European Conference on Computer Vision*. Springer, 2012, pp. 702–715.

[38] H. Possegger, T. Mauthner, and H. Bischof, "In defense of color-based model-free tracking," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2113–2120.

[39] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "Eco: Efficient convolution operators for tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6638–6646.
[40] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg, "Fast accurate scale space tracking," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2015, pp. 65.1–65.11.
[41] P. Dendorfer, A. Ošep, A. Milan, K. Schindler, D. Cremers, I. Reid, S. Roth, and L. Leal-Taixé, "MOTChallenge: A benchmark for single-camera multiple target tracking," *Int. J. Comput. Vis.*, vol. 129, pp. 845–881, 2021.

Jianguo Zhang is an Associate Research Fellow at the Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRS), Shenzhen Pengcheng Peacock Talent, Visiting Research Fellow at The Chinese University of Hong Kong, Shenzhen, and an Expert Member of the EPTC Power Robot Expert Committee. He serves as the R&D Director of the Robot Systems Center at AIRS, R&D Technical Lead of the Guangdong Embodied Intelligence Innovation Center, and Deputy Director of the Guangdong Embodied Intelligence Engineering Center. His current research focuses on digital twins, embodied intelligence technology, and multimodal large models. In 2024, he received the Second Prize of the Wu Wenjun Artificial Intelligence Science and Technology Progress Award from the Chinese Association for Artificial Intelligence, and in 2025, he was honored with the Second Prize of the Science and Technology Progress Award from the China Instrument and Control Society. He has years of experience in embodied intelligent robots, robot embodied intelligence perception and modeling, and embodied intelligence data acquisition.

Yun Chen was born in September 1993 in Jieyang, Guangdong Province. He holds a postgraduate degree and a master's degree in Control Science and Engineering. As a senior engineer, he currently serves as an expert at the Electric Power Research Institute of Guangdong Power Grid Co., Ltd.

Wenxing Sun was born in Guizhou in 1987. He graduated from Huazhong University of Science and Technology in 2017, majoring in High Voltage Technology, and holds a doctoral degree. His main research areas are high-voltage test diagnosis technology and equipment development for power equipment, as well as intelligent operation and maintenance technology for substations. At present, he serves as a senior manager in the Asset Section of the Production Technology Department of Guangdong Power Grid. He is the deputy secretary-general and member of the Power Robot Standards Committee in the energy industry, and a member of the IEC TC10 working group. He is also the secretary-general of the Subcommittee on Intelligent Inspection Technology for Power Transmission and Distribution (China) of IEEE and PES.

Yuhui Chen was born in June 1999 in Guangzhou, Guangdong Province. She has an undergraduate degree and a bachelor's degree in Electronic Science and Technology. As an assistant engineer, she is currently the deputy-duty engineer at the 500-kV Yuezhong Converter Station of the Digital Operation Department, the First Substation Management Institute of Guangzhou Power Supply Bureau, Guangdong Power Grid Co., Ltd.

Junwen Yao was born in Guangdong in 1997. She graduated from South China University of Technology in 2019, majoring in Electrical Engineering and Automation, with a bachelor's degree. Her main research field is substation intelligent technology. Currently, she serves as a specialist at the Mechanical and Electrical Research Institute of Guangdong Power Grid. She has work experience in the operation, maintenance and management of substation intelligent terminals, the inspection of machine-based patrol equipment, and the training of drone pilots. At present, she is mainly responsible for the management of substation-specialized intelligent terminals and participates in key scientific and technological projects of the provincial power grid company.

Hua Ye was born in May 1991 in Meizhou, Guangdong Province. He holds a graduate degree with a master's degree in Mechanical Engineering. He currently serves as a Senior Engineer at the Shenzhen Institute of Artificial Intelligence and Robotics for Society.

Duanjiao Li was born in Hunan Province in 1971. She holds a master's degree and graduated from Huazhong University of Science and Technology in 1993, majoring in High-Voltage Technology and Equipment. Her research focuses on substation intelligent technology. Currently, she serves as the General Manager of the Production Technology Department of Guangdong Power Grid. In addition, she is a strategic technical expert of China Southern Power Grid and the technical lead of the key R&D team for intelligent power transmission and transformation within China Southern Power Grid. In recent years, her achievements include five first-class, eight second-class, and seven third-class awards for scientific and technological progress of China Southern Power Grid; one first-class and one second-class award for value creation of China Southern Power Grid; one first-class and one second-class award for technical improvement contribution of China Southern Power Grid; one second-class award of the Guangdong Provincial Science and Technology Award; and one second-class award of the China Power Science and Technology Progress Award.