JOURNAL OF INFORMATION SCIENCE AND ENGINEERING XX, XXX-XXX (2026)
DOI:

A Dynamic Keyframe Extraction Strategy for Efficient and Accurate Person Tracking*

FENG-CHENG LIN+, HUAN LIANG ELVIN LEE
Department of Information Engineering and Computer Science
Feng Chia University
Taichung, 407102, Taiwan
E-mail: fclin@fcu.edu.tw;huanlianglee@gmail.com

A critical bottleneck persists in Multi-Object Tracking (MOT): the vast amount of redundant and low-quality data from state-of-the-art trackers creates a 'garbage in, garbage out' problem for downstream tasks. To address this, we propose an effective architectural alternative that focuses on intelligently refining the tracker's output rather than modifying the core tracking algorithm. Specifically, we introduce a novel Dynamic Four-Stage Keyframe Extraction strategy, embedded within the BoT-SORT tracker, which acts as an intelligent data purifier. By analyzing the tracker's internal state (confidence, spatial-temporal stability, and information entropy), our strategy selects a representative subset of frames for an efficient offline optimization stage. This approach reduces the computational load for downstream tasks like person re-identification by over 96.8% while dramatically improving key identity consistency metrics such as the IDF1 score, particularly in crowded scenes where baseline models fail. These results validate that tracker-aware data purification effectively decouples accuracy from processing volume, offering a pragmatic solution for deploying efficient, high-fidelity tracking systems in resource-constrained environments.

Keywords: Person Tracking, Multi-Object Tracking (MOT), Keyframe Extraction, Efficient Tracking, Data Reduction, Person Detection, BoT-SORT.

1. INTRODUCTION

Multi-Object Tracking (MOT) is a fundamental task in computer vision, aiming to locate all objects of interest within a video sequence and maintain their respective identities over time [1], [2]. The "tracking-by-detection" paradigm currently dominates the field, achieving state-of-the-art performance on authoritative benchmarks like the MOTChallenge [3], [4]. However, this success is built on a costly premise: to maintain maximum temporal accuracy, these methods must process every single frame. This frame-by-frame philosophy creates a fundamental efficiency bottleneck due to massive data redundancy [5]. The burden manifests across the entire system pipeline, from the heavy computational load for downstream tasks like person re-identification [6], to the increased real-time cost and memory pressure in integrated models [7]. Consequently, data redundancy has become a core obstacle to deploying efficient and scalable MOT systems in real-world applications [5], [8], [9], [10].

To combat data redundancy, several approaches have been explored [8], [11], [12], [13]. Naive fixed-rate subsampling is content-agnostic and risks missing critical events or producing sparse, irrelevant selections that disrupt temporal continuity [14]. More sophisticated methods propose adaptive keyframe selection based on image-level similarity

+ Corresponding author.

metrics like optical flow or learned feature distance [15]. While valuable, these tracker-agnostic strategies primarily answer a superficial question: “If the image looks different?” They fail to address the more critical, tracker-centric question: “If the state of the tracker itself is stable?” Minor visual changes might not be significant for identity, while subtle shifts in tracker confidence could signal an impending ID switch. This reveals a distinct research gap: the need for a filtering mechanism that leverages the tracker's rich internal state as a more intelligent signal for identifying truly critical frames [4]. Such a strategy moves beyond simple image analysis to pinpoint moments that are truly critical for maintaining identity, thereby addressing the “garbage in, garbage out” problem at its source [16].

To fill this gap, our research shifts the paradigm from improving the tracker to intelligently refining its output. We introduce a novel Dynamic Four-Stage Keyframe Extraction Strategy, which is embedded directly into the BoT-SORT [3] tracker and acts as an intelligent data "purifier". It integrates multi-dimensional, tracker-aware metrics, including detection confidence, spatial-temporal IoU changes, an anti-redundancy filter based on historical stability, and information entropy to select a sparse yet high-quality set of keyframes. To provide a stable and accurate foundation, we first systematically optimized a person detector, YOLO11m-CSA, which integrates Coordinate Spatial Attention [17] into the YOLO11 [18] architecture. Our design choices were validated through extensive ablation studies, as detailed in the methodology section. The main contributions of this article are summarized as follows.

We propose an online, four-stage keyframe extraction strategy that is tracker-aware, using the internal state of the tracker to guide the filtering process. Our method achieves over 96.8% data reduction on standard benchmarks while simultaneously enhancing key identity consistency metrics, such as the IDF1 score, particularly in the challenging crowded scenes.

We also demonstrate a highly adaptable strategy for enhancing the efficiency of multi-object tracking (MOT), showing that intelligently filtering the tracker's output is an effective and generalizable alternative to modifying the underlying tracking algorithm.

2. RELATED WORK

2.1 Multi-Object Tracking Paradigms

Multi-Object Tracking (MOT) aims to generate the trajectories of multiple objects of interest within a video sequence. The dominant and most successful approach in recent years is the "tracking-by-detection" paradigm [19]. This paradigm bifurcates the MOT problem into two distinct stages: a detection stage, where an object detector identifies all potential targets in each frame, and an association stage, where an algorithm links these detections across frames to form coherent trajectories. Early and foundational association methods often relied on the Kalman Filter to predict object motion and the Hungarian algorithm to resolve assignments based on spatial proximity [20]. More recent state-of-the-art trackers, such as BoT-SORT [3] and ByteTrack [4], have significantly improved association logic by incorporating appearance features from Re-ID models and more robustly handling occlusions and low-confidence detections. While these methods have pushed the boundaries of tracking accuracy, their reliance on

processing every frame creates the efficiency challenges we addressed in this work.

2.2 Efficiency and Data Reduction in MOT

The pursuit of higher accuracy in modern MOT systems has led to a significant bottleneck in efficiency and scalability. The most direct approach to mitigate data redundancy is **fixed-rate subsampling**, where frames are processed at a fixed interval. However, this content-agnostic strategy is inherently unreliable, as it may arbitrarily discard frames containing critical events, such as brief appearances or key interactions, leading to a degradation in tracking performance [5]. A second category of methods involves **offline video summarization**. While these techniques are effective at creating concise highlight reels for human consumption, they are fundamentally unsuited for real-time MOT systems, as their offline nature cannot support online decision-making, and their objective, exemplified by techniques like video synopsis [21], is not optimized for downstream algorithmic consumption.

More sophisticatedly, recent work by Yoon *et al.* [15] has proposed **Adaptive Frame Sampling (AFS)** based on image-level similarity metrics like optical flow or feature distance. However, we identify a key limitation: these strategies are **tracker-agnostic**. They primarily answer the question, **"Is the image different?"** not the more critical, tracker-centric question, **"Is the tracker's state stable?"** An alternative line of research focuses on **model-level efficiency**, for instance, by alternating between complex and simple models on different frames [22]. This approach addresses the per-frame computational cost but does not solve the underlying data redundancy problem faced by downstream tasks.

2.3 Pedestrian Detection in Crowded Scenes

The performance of the "tracking-by-detection" paradigm is fundamentally tethered to the quality of the initial detections [20]. Detecting pedestrians in crowded scenes poses a unique set of challenges, including heavy occlusion and significant scale variations, where distant individuals may only occupy a few pixels [23]. To combat these challenges, a significant body of research has focused on enhancing detector architecture. A particularly effective line of work involves the integration of **attention mechanisms**, which enable the network to selectively focus on more informative regions of the feature map [24]. Modules like the Convolutional Block Attention Module (CBAM) [25] and, more recently, Coordinate Attention [17] have proven effective. Coordinate Attention, in particular, excels at capturing long-range dependencies and preserving precise positional information, making it well-suited for identifying small, distant objects, a concept validated in works like PC-YOLO11s [26]. In summary, our review of the literature indicates that while the tracking-by-detection paradigm has excelled in accuracy, a critical research gap exists in efficiently managing the resulting data redundancy for downstream tasks. Prevailing data reduction techniques are often offline, content-agnostic, or, even when adaptive, remain tracker-agnostic, failing to leverage the tracker's rich internal state for more intelligent filtering. To address this specific gap, this paper introduces a novel, tracker-aware dynamic keyframe extraction strategy. Our approach is built upon a systematically optimized pedestrian detector to ensure a high-quality input foundation and is designed to intelligently refine the data stream from within the tracking process itself, offering a new path towards achieving both efficiency and accuracy in person tracking.

3. METHODOLOGY

In this section, we present our proposed framework, designed to achieve efficient and accurate person tracking by intelligently refining the data stream from a state-of-the-art tracker. The overall pipeline of our approach is illustrated in Fig. 1. We begin by providing a high-level overview of the system architecture. We then detail the core of our contribution: the Dynamic Four-Stage Keyframe Extraction Strategy, including the specific mechanisms of each filtering stage.

3.1 Overall Architecture

The proposed method enhances a standard tracking-by-detection framework by embedding a dynamic keyframe extraction strategy to improve identity consistency. The pipeline begins with an input video stream, where each frame is processed by a systematically optimized YOLO11m-CSA detector to generate robust pedestrian detections. These detections are then passed to the BoT-SORT tracker [3], which performs data association to establish and maintain initial, dense, frame-by-frame tracking results for each object.

Our core contribution is the **Dynamic Four-Stage Keyframe Extraction Strategy**, which is embedded within the BoT-SORT tracker to operate on initial tracking data in real-time. As the tracker processes each frame, this module acts as an intelligent data purifier, analyzing each active tracklet to select a sparse yet high-quality collection of keyframes. This curated set is optimized for efficient offline analysis, like person re-identification, to determine globally consistent identities. Crucially, these corrected identities are then applied back to the original tracking data, producing the final output for evaluation and ensuring a fair, protocol-aligned comparison.

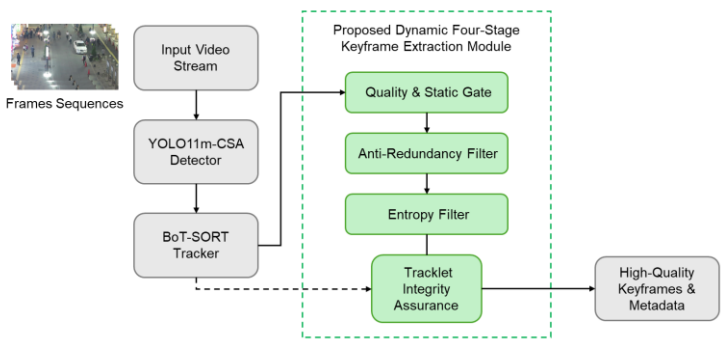


Fig. 1. The architecture of our proposed tracker-aware keyframe extraction strategy. After initial detection and tracking, our module analyzes each tracklet using a primary filtering cascade (solid line) and a parallel integrity assurance mechanism (dashed line), outputting a sparse, high-quality set of keyframes and metadata.

The internal architecture of this keyframe extraction module is illustrated in Fig. 1. It consists of two logical paths that operate on each tracklet: a primary filtering cascade (solid line) and a parallel integrity assurance mechanism (dashed line). The primary path is a deterministic cascade of three filters designed to assess the value of each frame: a **Quality & Static Gate**, an

Anti-Redundancy Filter, and an **Entropy Filter**. Concurrently, the secondary path implements a **Tracklet Integrity Assurance** mechanism, which acts as a fallback rule to guarantee that at least one representative frame is preserved for every valid tracklet.

3.2 YOLO11m-CSA Detector

As established in Section 3.1, our proposed dynamic keyframe extraction strategy is "tracker-aware," meaning its effectiveness relies on analyzing the state of the tracker's output. The quality of this state, however, is entirely dependent on the performance of the upstream detector. In crowded and multi-scale real-world scenarios, standard object detectors often struggle to reliably detect occluded or distant pedestrians, leading to unstable bounding boxes and fluctuating confidence scores [27], [28]. This signal "noise" can be misinterpreted by our content-aware filtering strategy as a significant event, thus triggering erroneous keyframe selections. Therefore, to ensure the efficacy of our core contribution, a custom-designed detector capable of providing a high-quality and stable detection signal in these complex scenes is an indispensable prerequisite.

Our design is primarily inspired by PC-YOLO11s, proposed by Wang et al. [26], a work that successfully enhanced performance on generic small-object detection tasks. Their approach involved two key architectural modifications: first, they replaced the original detection heads [P3, P4, P5] with [P2, P3, P4] to better leverage high-resolution feature maps for small targets, while removing the P5 head to improve efficiency. Second, they integrated the Coordinate Spatial Attention (CSA) module [17] to focus on the most salient feature regions. Inspired by this, we adapt and specialize this concept for the more challenging scenario of pedestrian detection. We developed our custom YOLO11m-CSA detector, whose detailed network architecture is illustrated in Fig. 2.

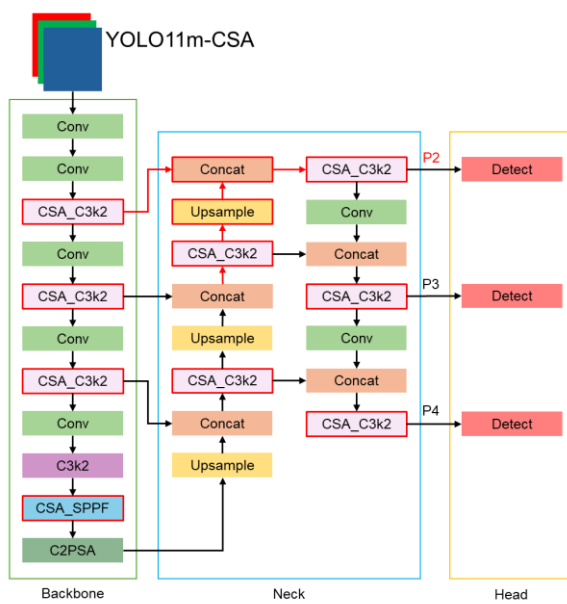


Fig. 2. The architecture of our proposed YOLO11m-CSA detector.

While inspired by PC-YOLO11s [26], our architecture is not a direct replication; instead, our contribution lies in the systematic optimization to find the most effective configuration for the person tracking task. Our specific architectural modifications include: First, inserting a lightweight CSA_C3k2 block at the deepest P6 layer to enhance high-level features without significant overhead. Second, replacing the standard SPPF module with a CSA-enhanced version (CSA_SPPF). Third, enabling residual connections at specific layers in the neck (P13 and P16) to preserve feature details during fusion.

The **Coordinate Spatial Attention (CSA)** Module. For clarity, we detail the internal mechanism of the CSA module [26], which enhances feature representation by embedding positional information. As illustrated in Fig. 3, the process is as follows: Let the input feature map be defined as $F \in \mathbb{R}^{C \times H \times W}$, where C, H, and W represent the number of channels, height, and width. First, this input map F is passed through a standard Spatial Attention Module (SAM) [25] to apply a preliminary weighting, resulting in a spatially-weighted feature map, F_{SAM}. Next, to capture long-range spatial dependencies, CSA generates coordinate information by applying max-pooling to F_{SAM} along the horizontal and vertical axes separately. We specifically utilize max-pooling, as it excels at capturing the most salient features of distant or small targets by preserving their peak signal responses. These critical signals might otherwise be diluted or lost entirely by the averaging effect of avg-pooling, especially amidst background noise. This produces two distinct direction-aware feature maps, $F_h \in \mathbb{R}^{C \times \mathrm{H} \times 1}$ (Equation 1) and $F_w \in \mathbb{R}^{C \times 1 \times W}$ (Equation 2).

$$F_h = AdaptiveMaxPool2d\left(F_{SAM,}(1, W) \right) \tag{1}$$

$$Fw = AdaptiveMaxPool2d\left(F_{SAM,}(H, 1) \right) \tag{2}$$

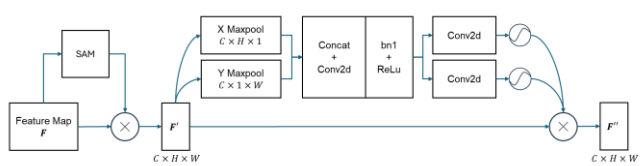


Fig. 3. The architecture of the Coordinate Spatial Attention (CSA) module, adapted from Wang *et al.* [26].

These two feature maps are then concatenated along the spatial dimension to form F_{cat}. This combined feature map is passed through a shared convolutional block, which includes Batch Normalization (BN) and a ReLU activation function, to fuse the information and reduce dimensionality, resulting in F_{reduce} (Equation 3).

$$F_{reduce} = ReLU\left(BN\left(Conv_{1 \times 1}, \left(F_{cat} \right) \right) \right) \tag{3}$$

The resulting feature is split and transformed back to the original dimension via separate

convolutional layers to generate independent horizontal A_h and vertical A_w attention maps. These two maps are ultimately combined via element-wise addition, passed through a Sigmoid function σ to produce the final coordinate attention map, A_{coord} (Equation 4). Finally, the original input feature map F is recalibrated by applying these weights through **element-wise mutiplication** (denoted by $\odot$), yielding the final, enhanced feature map F″ (Equation 5).

$$A_{coord} = \sigma\left(A_h^{II} + A_w^{W}\right) \tag{4}$$

$$F'' = F \odot A_{coord} \tag{5}$$

In summary, our final YOLO11m-CSA architecture is a synthesis of the design decisions detailed above. Crucially, each architectural choice, from the strategic placement of a lightweight CSA module to the activation of residual connections, was not arbitrary. Instead, each modification was systematically validated through an exhaustive set of ablation studies. The detailed performance metrics and quantitative results of these studies are presented in Section 4 (see Table 4). This rigorous optimization process yielded a foundational detector capable of providing stable, high-quality signals that are essential for the success of our subsequent tracker-aware extraction strategy.

3.3 Dynamic Four-Stage Keyframe Extraction Strategy

This section details the core contribution of our paper. Our design philosophy is to be tracker-aware, aiming to intelligently refine the data stream by utilizing the internal state of the tracker, rather than mere image content. To this end, we introduce a dynamic four-stage keyframe extraction strategy, implemented as a module that is seamlessly embedded within each tracklet instance of the BoT-SORT [3] tracker.

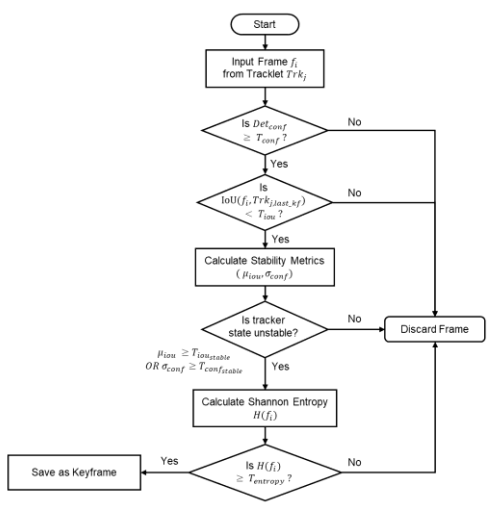


Fig. 4. Decision flowchart of the proposed keyframe extraction method.

Within this tracker-aware framework, we first define an ideal keyframe by three core properties: 1) **high fidelity**, originating from a high-confidence detection; 2) **high information value**, capturing a significant change in the state of the target; and 3) **high-quality sample**, featuring clear and non-occluded visual content. Our proposed four-stage filtering cascade is designed to select frames that best approximate these ideal properties using a series of computable metrics. The detailed decision-making process is illustrated in Fig. 4.

The strategy evaluates each tracklet using a deterministic three-stage filtering cascade: (A) a **Quality & Static Gate** for initial filtering; (B) an **Anti-Redundancy Filter**; (C) **Entropy-based Curation** to preserve information value and sample quality. To prevent catastrophic data loss, it executes (D) **Tracklet Integrity Assurance** as an overriding initial step (Stage 0) before the cascade. The logic of each stage is detailed below:

(A) *Quality & Static Gate*

This stage is the first line of defense in our filtering process, designed to quickly remove many unsuitable frames with low computational cost. This gate is based on two properties of an ideal keyframe. First, to ensure high fidelity, we follow a simple principle: a reliable keyframe should come from a detection where the model is highly confident. Low-confidence detections are often noisy and do not represent the target well. Second, to ensure high information value, we filter out frames where the target is stationary, as they provide no new information about the trajectory. This gate applies two criteria. For a given detection det_i with a confidence score c_i and bounding box $bbox_i$, it must first meet the condition $c_i \geq T_{conf}$. Next, to filter static frames, the Intersection over Union (IoU) between its bounding box b_i and its bounding box from the previous frame, $Trk_j.bbox_{prev}$, must satisfy $IoU(bbox_i, Trk_j.bbox_{prev}) < T_{iou}$ to ensure the target has moved significantly.

(B) *Anti-Redundancy Filter*

After a frame passes the initial gate, this more advanced filter ensures that any captured motion is genuine and significant, not just noise from the tracking process. While the integration of the Coordinate Spatial Attention (CSA) module significantly enhances the detector's spatial focus, single-frame object detection remains inherently vulnerable to "sophisticated garbage"—such as highly detailed background objects or complex occlusions that mimic human features and yield deceptively high confidence scores.

To mitigate this vulnerability, our Anti-Redundancy Filter shifts the filtering paradigm from pure spatial analysis to temporal kinematics. The rationale is that "sophisticated garbage" (static background artifacts) or smoothly moving targets will show an unnatural, rigid stability in IoU and confidence trends over a short period, whereas a significant real-world event will disrupt this stability. To achieve this, our mechanism analyzes historical data trends within a time window of W_{hist} frames. A frame is considered redundant and discarded only if two conditions are met simultaneously. First, the absolute deviation of the current IoU from the mean of its history $\mu_{iou_{his}}$, is less than a stability threshold T_{iou_stable}. Second, the standard deviation of the confidence history σ_{conf}, is less than a stability thresholdT_{conf_stable}. This double-check mechanism leverages temporal gating to effectively discard deceptive high-quality noise while remaining highly sensitive to significant dynamic changes.

(C) *Entropy-based Curation*

This final stage in our filtering cascade acts as a quality control step to ensure that each selected keyframe is a high-quality visual sample. The motivation for this step is that a frame might capture a genuine motion event but be of poor visual quality due to motion blur or heavy occlusion. Such frames do not satisfy the high-quality sample property, as they would yield noisy and unreliable feature vectors from a downstream Re-ID model. Therefore, a mechanism is needed to quantify the visual information content within the bounding box.

To achieve this, our mechanism uses Shannon Entropy [29] as a proxy for image quality. To improve efficiency, this calculation is only triggered if the current frame has passed a minimum gap threshold, T_{gap}, from the last keyframe. The process is as follows: first, the region defined by the bounding box $bbox_i$ is cropped. The crop is then converted to a grayscale image to calculate its pixel intensity histogram. Finally, Shannon Entropy is computed from this histogram, as defined by:

$$H(X) = -\sum_i p(x_i) \log_2 p(x_i) \text{ [29]} \tag{6}$$

Where *H(X)* is the Shannon Entropy of the image region, *X* is the random variable representing the pixel intensity values, x_i is a possible intensity value in grayscale levels, and $p(x_i)$ is the probability of that pixel value occurring in the region, calculated from the normalized histogram. Frames with an entropy value below a threshold $T_{entropy}$ are discarded, as low entropy typically indicates motion blur or a lack of visual detail.

(D) *Tracklet Integrity Assurance*

This component operates as a parallel safeguard mechanism rather than a sequential filter. Evaluated as the highest-priority rule (Stage 0) in our algorithm, its purpose is to ensure the completeness and robustness of the final keyframe set. The motivation for this is that the three-stage filtering cascade is intentionally strict in its pursuit of high-quality samples. However, this strictness can lead to an edge case where a valid but brief tracklet has all its frames discarded. This would result in the complete loss of that identity for any downstream analysis, which constitutes a major system failure.

To prevent this, our mechanism introduces a key safeguard rule to ensure the completeness of the final keyframe set. This rule addresses the edge case where a valid tracklet might have all its frames discarded by the strict filtering cascade. The mechanism stipulates that for any new tracklet, an initial keyframe is forcibly saved once its age, $Trk_j.age$, reaches a minimum stabilization period, T_{min_age}. This ensures that all newly established, stable tracks are represented by at least one visual sample, preventing their complete loss from downstream analysis. To formally summarize the logic of the entire strategy, we present it as the following pseudocode in Algorithm 1.

Algorithm 1 Tracker-Aware Keyframe Extraction for a Single Tracklet

Input: Current tracklet state Trk_j; new detection det_i with (confidence c_i and bounding box $bbox_i$); current frame ID f_{id} full frame image Img_i

Output: A boolean flag $is_keyframe$; updated $Trk_j.keyframes$ list

Initialize: $is_keyframe \leftarrow False$;

Update $Trk_j.iou_history, Trk_j.conf_history$ with information from det_i

/* Stage 0: Tracklet Integrity Assurance */

if Trk_j has no keyframe and $Trk_j.age \geq T_{min_age}$ **then**

  $SaveKeyframe(Trk_j, det_i, f_{id})$

  return $True$

end if

/* Stage 1: Quality & Static Gate */

if $c_i < T_{conf}$ or $IoU(bbox_i, Trk_j.bbox_{prev}) > T_{iou}$ **then**

  return $False$

end if

/* Stage 2: Anti-Redundancy Filter */

$\mu_{iou} \leftarrow Mean(Trk_j.iou_history)$

$\sigma_{conf} \leftarrow StandardDeviation(Trk_j.conf_history)$

if $abs(Trk_j.iou_history.last - \mu_{iou}) < T_{iou_{stable}}$ and $\sigma_{conf} < T_{conf_{stable}}$ **then**

/* Frame is considered redundant if both motion and confidence are stable */

  return $False$

end if

/* Stage 3: Entropy-based Curation */

if $(f_{id} - Trk_j.lastkeyframe_id) \geq T_{gap}$ **then**

  return $False$

end if

$entropy \leftarrow ShannonEntropy(Img_i, bbox_i)$

if $entropy < T_{entropy}$ **then**

  return $False$

end if

$SaveKeyframe(Trk_j, det_i, f_{id})$

return True

4. EXPERIMENTS

This section presents a series of comprehensive experiments designed to validate our proposed keyframe extraction strategy. We evaluate its performance from three primary aspects: 1) the contribution of each component via ablation studies, 2) the overall tracking accuracy and data reduction performance, and 3) the impact on downstream clustering quality.

4.1 Dataset and Evaluation Protocol

We evaluated our proposed method on a series of public datasets covering detector training, tracking performance evaluation, and downstream task validation. These datasets include CrowdHuman [27], WiderPerson [28], the MOTChallenge [30], EPFL [31], Market-1501 [32], and DukeMTMC-reID [33].

(A) CrowdHuman

This dataset focuses on high-density crowd scenarios, providing a foundational training set for our detector to handle severe occlusions. To ensure annotation consistency, we utilized only the "Full Bounding-Box" for training the model. This dataset contains 15,000 images for training, 4,370 images for validation, and 5,000 for testing.

(B) WiderPerson

The WiderPerson dataset enhances our model's generalization ability by providing a large corpus of images from diverse scenarios. We merged its original "pedestrians" and "riders" categories into a single "person" class and similarly used only full-body annotations. This dataset contains 8,000 training images, 1,000 validation images, and 4,382 testing images. We combined it with CrowdHuman to construct the final training set for our detector.

(C) MOTChallenge

This benchmark serves as our primary platform for evaluating core tracking performance and efficiency. We selected a challenging subset of sequences from MOT17 and MOT20 to form a rigorous and diverse testbed, with their properties detailed in Table 1.

Table 1
MOTChallenge Dataset configuration

Scenario	Density	Duration	Boxes	Frame Rate	Description
MOT17-04	45.3	½ min	47,557	30 fps	Pedestrian street at night, elevated viewpoint.
MOT17-09	10.1	½ min	5,325	30 fps	A pedestrian street scene filmed from a low angle.
MOT20-02	72.7	2 min	202,215	25 fps	Crowded indoor train station.
MOT20-05	226.6	2 min	751,330	25 fps	Crowded square by night time.

(D) EPFL

The EPFL dataset is employed to evaluate the effectiveness of our extracted keyframes for downstream, cross-camera analysis tasks. It provides two representative scenarios, “Laboratory” (indoor) and “Terrace” (outdoor), with their configurations detailed in Table 2. Both scenarios feature footage from 4 synchronized cameras with annotated bounding boxes and consistent cross-camera track IDs, making the dataset an ideal platform for rigorously evaluating our method’s performance in a multi-camera setting.

Table 2
EPFL Dataset configuration

Scenario	Cameras	Density	Duration	Frame Rate
Laboratory	4	6.7	2½ min	25fps
Terrace	4	10.8	3½ min	25fps

(E) Market-1501

This dataset consists of 32,668 annotated images of 1,501 identities captured by 6 cameras, used to train the appearance feature extractor of re-ID. The training set contains 19,732 images of 751 identities, while the test set is composed of 3,368 query images and 17,661 gallery images of 750 identities.

(F) DukeMTMC-reID

This large-scale benchmark was captured by 8 synchronized cameras and contains 36,411 images of 1,404 identities. Its training set contains 16,522 images of 702 identities, while its test set includes 2,228 query images and 17,661 gallery images.

We evaluate the performance of the model using a comprehensive set of metrics. For object detection performance, we use precision and recall, measuring the accuracy and completeness of the detections, and mean Average Precision (mAP@0.5) as the overall performance indicator at an IoU threshold of 0.5. For tracking accuracy, we adopt the standard MOT metrics [2], including Multiple Object Tracking Accuracy (MOTA), which aggregates errors in the tracking process; Identification F1-Score (IDF1), which focuses on identity consistency; and the number of ID Switches (IDSW). For efficiency, we quantify performance using the Data Reduction Ratio. For the downstream task quality, we use Purity.

To validate our core hypothesis that the keyframe strategy mitigates the "garbage in, garbage out" (GIGO) problem, we designed and compared two distinct offline optimization pipelines. Critically, both pipelines begin with the **same initial tracking data** and produce a final, **frame-by-frame output**. This ensures our comparison is strictly aligned with the MOTChallenge protocol. The two configurations are defined as follows:

(A) Naive Full-Frame Clustering

This configuration serves as a diagnostic control experiment to explicitly quantify the GIGO problem. In this pipeline, the initial tracking output is processed by feeding **all frames** from every tracklet into the Re-ID module for feature extraction and global clustering. The resulting (and often erroneous) cluster information then updates the initial track IDs. This setup simulates a direct but flawed approach that disregards data quality.

(B) Keyframe-Guided Purified Clustering

We apply our four-stage keyframe strategy to the initial tracking data to select a high-quality, sparse subset of frames. Only these purified keyframes undergo Re-ID feature extraction and clustering. The resulting superior cluster information then updates the initial, dense track IDs to produce the final, high-accuracy output.

As will be shown in Section 4.4, the dramatic performance difference between these two configurations does not stem from any misalignment in the evaluation protocol. Instead, this difference directly measures the efficacy of our keyframe strategy as a data purifier and validates its ability to solve the GIGO problem at its source.

4.2 Implementation Details

All experiments were conducted using the PyTorch framework with CUDA 12.4. The detection and tracking experiments were performed on a server equipped with an NVIDIA GeForce RTX 3090 GPU, while the Re-ID model for downstream validation was trained on a platform with an NVIDIA GeForce RTX 4070 Ti GPU and an Intel Core i9-13900K CPU.

Our detector was developed in two stages. First, we benchmarked mainstream YOLO models on the CrowdHuman dataset to select **YOLO11m** as the baseline. Second, our final YOLO11m-CSA model was trained and evaluated on a combined dataset of CrowdHuman and WiderPerson. For all detector training, we used an input resolution of 800x800, a batch size of 8, the SGD optimizer with a learning rate of 1e-3, and Mosaic data augmentation, and were trained for 300 epochs.

For the tracking stage, we utilized the official implementation of the **BoT-SORT** [3] tracker, retaining most of its default hyperparameters for data association. Notably, to focus specifically on evaluating our tracker-state-based filtering strategy, the internal Re-ID feature of BoT-SORT [3] was not used for association in this stage. The hyperparameters for our keyframe extraction strategy were set empirically based on performance on a validation set. For the Quality & Static Gate, the confidence threshold T_{conf} was set to 0.75 and the static IoU threshold T_{iou} was 0.90. The Anti-Redundancy Filter operated over a 10-frame window W_{hist} with stability thresholds for IoU T_{iou_stable} and confidence T_{conf_stable} set to 0.05 and 0.08. The final entropy threshold, $T_{entropy}$ was 3.4. For Tracklet Integrity Assurance, the minimum tracklet age T_{min_age} and minimum keyframe gap T_{gap} were set to 5 and 25 frames, respectively.

The appearance feature extractor used for downstream task validation employed a **ResNet-50** backbone. This model was trained on the Market-1501 [32] dataset using the Group

Sampling strategy [34]. During training, input images were resized to 256x128, and the model was trained for 50 epochs with a batch size of 16. We used the SGD optimizer with a momentum of 0.9 and an initial learning rate of 3.5e-4, which was decayed by a factor of 10 every 20 epochs. Data augmentation techniques, including random horizontal flipping and random erasing, were also applied. On the Market-1501 test set, this model achieved 91.8% Rank-1 accuracy and 81.5% mAP, confirming its effectiveness as a strong feature extractor for our validation experiments.

4.3 Ablation Studies

Baseline Detector Selection: To select an optimal baseline model that balances accuracy and efficiency for our subsequent optimizations, we first conducted a comprehensive performance evaluation of mainstream YOLO models (v5 [35], v8 [36], v9 [37], v10 [38], and YOLO11 [18]) on the CrowdHuman [27] dataset. The evaluation focused on two primary criteria: detection accuracy, measured by mean Average Precision (mAP), and model efficiency, measured by inference speed in PyTorch. The results are visualized in Fig. 5, which plots mAP against inference speed, where the top-left quadrant represents the most desirable performance. The plot reveals a clear "efficiency frontier" on which the YOLO11 and YOLOv10s series models are the most competitive. For a definitive selection, Table 3 details the performance of the YOLO11 series. We ultimately selected **YOLO11m** as our baseline model. As shown, YOLO11m achieves a high mAP@0.5:0.95 of 63.0% while maintaining a fast inference speed of 3.331 ms and a moderate parameter count of 20M. Due to this superior balance between accuracy, speed, and model size, we chose it as the foundation for all subsequent optimizations and comparisons.

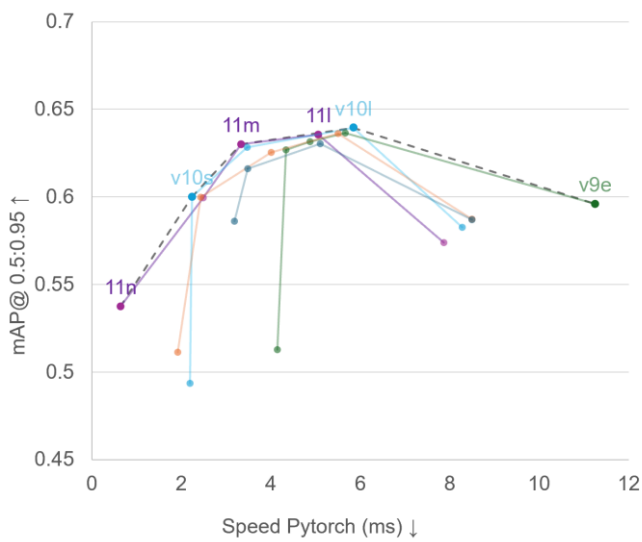


Fig. 5. Performance comparison of various YOLO models on the CrowdHuman dataset. The x-axis represents inference speed (lower is better), and the y-axis represents mAP@0.5:0.95 (higher is better).

Table 3.
Detailed performance of the YOLO11 series models. mAP@0.5 (%), mAP@0.5:0.95 (%), speed PyTorch (ms), and parameters (M) are reported.

Model	mAP@0.5	mAP@0.5:0.95	Speed PyTorch	Parameters
YOLO11n	83.6	53.8	**0.646**	**2.590**
YOLOv10s	85.2	60.0	2.239	8.067
YOLO11m	87.5	63.0	3.331	20.054
YOLO11l	**90.3**	**63.6**	5.062	25.311

Optimization of the YOLO11m: With YOLO11m established as our baseline, we conducted a series of ablation studies to determine the optimal integration strategy for the CSA module, with the results summarized in Table 4. Our initial experiment, inspired by the broad integration approach in PC-YOLO11s [26], involved applying CSA modules across multiple layers. This approach yielded significant performance gains, increasing the mAP@0.5 from 87.53% to 89.83% and boosting recall by over 2%. However, this gain came at a significant cost to efficiency, with the inference time increasing from 3.331 ms to 5.873 ms, motivating our search for a more strategic integration. To address the efficiency issue, we applied a lightweight CSA module (with a channel ratio of $e = 0.25$) only at the deepest P6 layer. As shown in Table 4, this approach successfully maintained a high accuracy with mAP@0.5 of 89.89%, while reducing the inference time to 5.096 ms. As a final refinement, we enabled residual connections at the P13 and P16 layers of the neck.

Table 4.
Ablation Study of the YOLO11m-CSA Architecture on the CrowdHuman Dataset. precision (%), recall (%), mAP@0.5 (%), speed PyTorch (ms), parameters (M), and GFLOPs are reported.

Model	Precision	Recall	mAP@0.5	Speed	Parameters	GFLOPs
YOLO11m (Baseline)	87.86	78.80	87.53	**3.331**	20.054	**68.5**
+ Broad CSA Integration	88.14	81.13	89.83	5.873	16.560	80.5
+ Strategic Lightweight CSA (P6)	88.23	81.44	89.89	5.096	**15.693**	77.8
+ Residual Connections (Ours)	**87.88**	**81.82**	**90.38**	5.018	**15.693**	77.8

Performance of the YOLO11m-CSA: To quantify the performance of our final detector, we conducted a head-to-head comparison between YOLO11m-CSA and its baseline, YOLO11m, on the combined CrowdHuman [27] and WiderPerson [28] datasets. As detailed in Table 5, our YOLO11m-CSA model achieves an mAP@0.5 of 92.62%, a significant 1.77% improvement over the baseline's 90.85%. To further investigate the source of this gain, we analyzed the COCO-style scale-based metrics (Table 6). The results indicate that the improvement is particularly pronounced for small objects, where recall increased by 4.7%, confirming that the CSA module enhances perception of distant pedestrians.

Table 5.
Performance comparison between YOLO11m-CSA and its baseline on the combined dataset. Precision (%), recall (%), mAP@0.5 (%), Speed PyTorch (ms), parameters (M), and GFLOPS are reported.

Model	Precision	Recall	mAP@0.5	Speed	Parameters	GFLOPs
YOLO11m	87.28	82.53	90.85	**4.516**	20.054	**68.5**
YOLO11m-CSA (ours)	**88.01**	**85.07**	**92.62**	6.246	**15.693**	77.8

Analysis of the Keyframe Extraction Strategy: To quantify the individual contribution of each component within our four-stage strategy, we conducted a detailed ablation study on the MOT20-01 sequence, with the results summarized in Table 7. In the following analysis, we use the original tracking results (no keyframe extraction) as the accuracy baseline and progressively add each of our filtering and safeguard mechanisms to demonstrate how our final strategy maintains and even improves tracking performance while achieving significant data reduction.

Table 6.
COCO-style scale-based metrics analysis for YOLO11m-CSA and the baseline model. Precision@0.5:0.95 (%), and recal@0.5:0.95 (%) are reported.

Model	Precision@0.5:0.95			Recall@0.5:0.95		
	Small	Medium	Large	Small	Medium	Large
YOLO11m	39.1	67.6	75.1	51.9	75.8	81.5
YOLO11m-CSA (ours)	**42.7**	**69.7**	**77.1**	**56.6**	**77.4**	**83.4**

As shown in Table 7, our analysis begins with the baseline performance without keyframe extraction: the system processes all 18,178 frames, achieving an IDF1 score of 71.6% with 144 ID switches. When applying only the initial **Quality & Static Gate**, the

frame count is drastically reduced to 544, an efficiency gain of over 96.8%. However, this aggressive filtering incurs a significant penalty to accuracy: the IDF1 score drops sharply to 65.8%, and the number of ID switches increases. This result demonstrates that a naive filtering approach, while efficient, inevitably discards critical information required for maintaining identity consistency.

Table 7.
Ablation study of the proposed keyframe extraction strategy on the MOT20-01 sequence. Frames, ID F1-score (%), and ID switch are reported.

Method Configuration	Frames	IDF1	IDSW
Baseline (No Keyframe Extraction)	18,178	71.6	**144**
+ Quality & Static Gate	544	65.8	149
+ Integrity Assurance	663	67.9	139
+ Anti-Redundancy Filter	594	70.0	135
+ Entropy Filter (ours)	**578**	**71.6**	**128**

To address this loss of critical information, our next configuration introduces the **Tracklet Integrity Assurance** mechanism. As shown in Table 7, this addition substantially recovers the IDF1 score from 65.8% to 67.9% and reduces ID switches from 149 to 139, at the cost of a slight increase in retained frames to 663. This demonstrates the crucial role of this safeguard. By ensuring at least one representative sample is kept for every valid tracklet, it successfully prevents the complete loss of identity information for brief tracks that would otherwise be entirely filtered out. However, the IDF1 score at this stage still lags the baseline, indicating that a more refined filtering stage is necessary to remove noise.

To further close the performance gap to the baseline, our next configuration incorporates the **Anti-Redundancy Filter**. As detailed in Table 7, this history-based filter provides a significant boost: the IDF1 score improves further to 70.0%, and ID switches are reduced to 135. Notably, this gain in accuracy is achieved while simultaneously reducing the number of keyframes from 663 to 594. This result validates the effectiveness of our tracker-aware stability check. By filtering out frames where the target is only experiencing minor jitter, this stage removes noisy, redundant data that can confuse the identity association process, leading to cleaner and more consistent tracklets.

Finally, we incorporated the **Entropy-based Curation** stage to ensure the visual quality of the final samples. As shown in Table 7, this last filter restores the IDF1 score to the baseline level of 71.6% and further reduces ID switches to a minimum of 128, all while processing only 578 frames. This ablation study demonstrates the effectiveness of our comprehensive strategy: each component plays an indispensable role, collectively achieving a data reduction of over 96% while maintaining or even improving key identity consistency metrics.

4.4 Main Tracking Performance

To validate the final performance of our complete framework, we conducted a direct comparison between our method, integrated with the keyframe extraction strategy and three settings on several challenging MOTChallenge [30] sequences: a naive no-filtering baseline, Fixed-Rate Subsampling (FRS), and Adaptive Frame Sampling (AFS). All experiments were conducted using our finalized YOLO11m-CSA detector and the BoT-SORT [3] tracker.

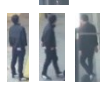
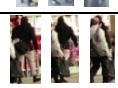
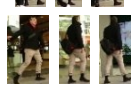
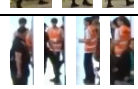
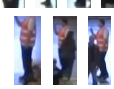
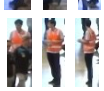
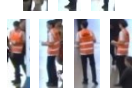
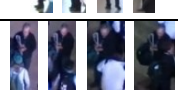
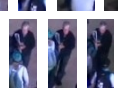
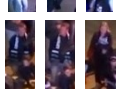
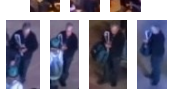
As detailed in Table 8, our method achieves a data reduction of over 96.8% on moderate-density MOT17 sequences while simultaneously enhancing the IDF1 score (e.g., from 61.7% to 73.6% on MOT17-04). The strategy's most profound impact is on crowded MOT20 sequences. It is crucial to note that our tracker-only pipeline (without Re-ID) achieves a high MOTA of 73.4% on MOT20-02. However, this setup lacks the embeddings needed for cross-camera tasks. The baseline's performance collapse (9.6% MOTA) occurs when Re-ID is naively applied to every frame.

To ensure a fair evaluation of our extraction strategy against existing subsampling methods under an identical data budget, FRS and AFS were configured to extract approximately the same number of bounding boxes as our method. The results show that FRS is content-agnostic and blindly discards critical tracking events, while global-level AFS is easily disrupted by irrelevant scene dynamics rather than target-specific variations. Consequently, both FRS and AFS fail to address the critical "garbage in, garbage out" problem at the feature extraction stage: feeding numerous occluded, blurry, and ambiguous crops to the downstream Re-ID model severely contaminates its feature vectors, causing catastrophic association failures that undermine the tracker's original accuracy. In contrast, our object-centric purification process restores the MOTA to 73.5% on MOT20-02. This improvement is visually substantiated in the final column of Table 8, which contrasts the low-quality, ambiguous frames processed by the baseline against the clean, representative keyframes selected by our method.

Table 8 transparently reports the identity switches (IDSW) for our method, which is labeled as "Ours" in the Method column. While IDSW increases in these highly compressed scenarios across all subsampling methods, this is a secondary effect of producing numerous high-purity, albeit shorter, tracklets where the baseline's associations would otherwise collapse completely. This makes the substantial gain in overall identity consistency (IDF1) a more meaningful indicator of success. These results suggest that tracker-aware filtering is not merely a tool for efficiency, but a crucial mechanism for enabling Re-ID-based downstream tasks without sacrificing the tracker's intrinsic accuracy.

We explicitly clarify that while IDSW evaluates *short-term* tracking continuity, IDF1 measures *global* identity preservation. Our framework generates shorter but higher-purity tracklets, which are then reconnected by our offline Re-ID clustering module. As demonstrated on sequence MOT20-02, despite the higher IDSW, our method achieves a substantial gain in global identity consistency, reaching an IDF1 of 63.5% (compared with the baseline's 19.2% and FRS's 51.9%). This confirms that the substantial gains in overall tracking accuracy justify the increase in short-term ID switches.

Table 8.
Tracking performance comparison of keyframe extraction strategies across multiple MOT sequences. The 'Baseline' represents a naive full-frame clustering approach designed to quantify the GIGO problem. Fixed-Rate Subsampling (FRS) and Adaptive Frame Sampling (AFS) [15] are configured to extract approximately the same number of bounding boxes as our proposed method. 'Ours' employs the tracker-aware keyframe strategy for data purification prior to ID correction. All methods are evaluated on representative continuous segments of the sequences to ensure a strictly fair comparison.

Sequence	Method	BBoxes for Re-ID (s)	MOTA (%)	IDF1 (%)	IDSW	Input Frames
MOT17-04	Baseline	12,346	42.7	61.7	20	
	FRS	265	45.5	64.3	20	
	AFS [15]	**260**	44.7	63.0	**16**	
	Ours	265	**65.8**	**73.6**	59	
MOT17-09	Baseline	3,622	53.0	58.1	18	
	FRS	155	50.7	62.6	17	
	AFS [15]	**147**	51.5	61.7	**16**	
	Ours	153	**54.0**	**62.7**	31	
MOT20-02	Baseline	10,420	9.6	19.2	**49**	
	FRS	2,129	70.9	51.9	584	
	AFS [15]	2,153	22.2	34.7	68	
	Ours	**2,164**	**73.5**	**63.5**	575	
MOT20-05	Baseline	48,947	11.9	22.3	176	
	FRS	**4,627**	11.0	17.5	180	
	AFS [15]	4,670	14.2	26.6	**125**	
	Ours	4,720	**62.9**	**63.0**	1,754	

To explicitly evaluate the generalizability of our extraction strategy and verify that it is not merely a tracker-specific optimization, we further integrated our module into the ByteTrack [4] framework. Since BoT-SORT [3] fundamentally inherits the core BYTE association logic, our extraction strategy can be seamlessly embedded into ByteTrack's update loop with only minimal adjustments.

As demonstrated in Table 9, we evaluated this integration on subsets of MOTChallenge representing both standard (MOT17-04) and extremely crowded (MOT20-02) scenarios. The results show that our extraction module is highly adaptable to ByteTrack, achieving data reduction rates of over 99%. Specifically, from an initial volume of 47,557 bounding boxes in MOT17-04, it selects only 215 highly representative keyframes, yielding a 99.5% reduction, while for the extremely crowded MOT20-02 sequence, it reduces 202,215 bounding boxes down to just 1,287 keyframes (a 99.4% reduction). This empirical evidence validates our approach as a generalizable, highly adaptable algorithmic extension for broader tracking-by-detection frameworks.

Table 9.
Compatibility and Data Reduction Performance of the Proposed Strategy Integrated with ByteTrack.

Sequence	Tracker	Original BBoxes	Extracted Keyframes	Data Reduction Rate
MOT17-04	ByteTrack	47,557	215	99.5%
MOT20-02	ByteTrack	202,215	1,287	99.4%

4.5 Computational Overhead Analysis for Real-Time Processing

To ensure the system's suitability for real-time deployment, we conducted a micro-benchmark analysis to quantify the additional processing load introduced by the Four-Stage Keyframe Extraction module. As summarized in Table 10, the extraction strategy is implemented as a highly efficient cascaded filter.

Table 10.
Micro-benchmark of the additional computational overhead introduced by the Four-Stage Keyframe Extraction module (tested on MOT17-04 sequence).

Stage	Processing Component	Execution Frequency	Avg. Latency (ms/box)	Hit(s)
Stage 1 & 2	Confidence & IoU checks	All candidate boxes	0.0404	41,000
Stage 3	Area Cropping (Slicing)	Filtered boxes only	0.0037	227
Stage 3	Shannon Entropy Calculation	Filtered boxes only	0.0657	227
Overall System	**Amortized Overhead**	**Per video frame**	**0.1098**	

The preliminary stages (confidence and bounding-box IoU checks) involve simple scalar operations and introduce a negligible latency of approximately 0.0404 ms per bounding

box. The more computationally intensive operations in the third stage involve isolating the region of interest via memory-efficient NumPy array slicing and calculating the Shannon Entropy. These operations take approximately 0.0037 ms and 0.0657 ms per candidate box, respectively.

These operations are not executed on every detected object. The MOT17-04 sequence consists of 1,050 video frames. During our experiments on this sequence, over 41,000 bounding box evaluations were processed by the preliminary stages. Only 227 candidate boxes (approximately 0.55%) triggered the final cropping and entropy calculations. A bounding box processed through the entire cascade incurs an average latency of 0.1098 ms. Given an average of 39.0 detections per frame (41,000 boxes across 1,050 frames in MOT17-04), this early-rejection mechanism yields an amortized overhead of only 1.59 ms per video frame. The complete end-to-end tracking pipeline, which includes video frame decoding, YOLO11m-CSA detection, Non-Maximum Suppression (NMS), and BoT-SORT tracker updates, requires an average processing time of 86.0 ms per frame. Therefore, our keyframe extraction module adds less than 2% to the total system latency and fully preserves the real-time tracking capability of the framework.

4.6 Impact on Downstream Task and Qualitative Analysis

To quantify the impact of keyframes on downstream tasks, we designed a cross-camera clustering experiment. This experiment utilized a two-stage process on the extracted Re-ID features: first, in a local stage, DBSCAN was applied to the keyframe features within each tracklet to filter outliers and form representative sub-clusters; second, on a global stage, the centroids of all sub-clusters were aggregated and clustered again to merge identities across cameras.

We use cluster Purity as the evaluation metric to compare the results of using all frames versus using only the keyframes extracted by our strategy. As shown in Table 11, across all test sequences, clustering only the sparse set of keyframes consistently resulted in higher Purity scores than clustering all frames. This strongly indicates that our strategy effectively filters out noisy and redundant samples that interfere with feature learning, yielding a cleaner, higher-quality dataset for downstream tasks.

Table 11.
Impact of keyframe extraction on downstream clustering purity. Clustering Purity (%) is reported.

Sequence	Clustering Purity	
	Using All Frames	Using Keyframes Only
MOT17-04	97.68	99.11
MOT17-09	94.39	96.18
MOT20-02	88.17	90.55
MOT20-05	95.08	97.82

Fig. 6 provides a qualitative illustration of the data reduction achieved by our method. For a single tracklet originally containing 3,122 frames, our strategy selected only 62 representative keyframes. This demonstrates a data reduction of over 98% while successfully preserving the core appearance changes of the target, significantly lowering the storage and computational costs for subsequent analysis.

TrackID	Number of Person Frames	Example Frames
1 (without extractor)	3122	
1 (with extractor)	62	
4 (without extractor)	1626	
4 (with extractor)	32	

Fig. 6. Qualitative illustration of keyframe extraction on a single tracklet from the EPFL dataset. Our method distills a long, redundant sequence into a compact set of representative frames.

4.7 Limitations and Failure Cases

While the proposed four-stage extraction strategy significantly purifies the data for downstream tasks, the reliance on Shannon Entropy as a final quality gate introduces certain statistical limitations. Shannon Entropy measures the dispersion of pixel intensities within a bounding box. Although it effectively filters out uniform, low-information crops, it does not always align perfectly with human visual perception of image quality.

Fig. 7 illustrates these specific statistical anomalies. As shown in Fig. 7(a), a crop suffering from severe motion blur and heavy background interference from another pedestrian in a high-contrast orange vest yields a misleadingly high entropy score (4.70). The mixture of blurred target edges and contrasting background colors artificially flattens the pixel intensity histogram, causing the system to accept this ambiguous crop. Fig. 7(b) demonstrates a perfectly clear, unoccluded target yields a low entropy score (3.10). The uniform texture of the background and the solid colors of the target's clothing mathematically concentrate the histogram, undervaluing a highly representative keyframe.



Fig. 7. Limitations of Shannon Entropy in quality assessment. (a) A misleading high-entropy frame. Severe motion blur and high-contrast background clutter artificially inflate the score. (b) A high-quality but low-entropy frame. Uniform clothing and background textures mathematically reduce the score despite the target being perfectly clear (selected images from MOT20-02).

Addressing this limitation presents a strict architectural trade-off. Deploying a dedicated deep learning-based Image Quality Assessment (IQA) model could effectively resolve these edge cases. However, such neural networks introduce substantial computational latency. As analyzed in Section 4.5, computing Shannon Entropy requires only 0.0657 ms per candidate box. This near-zero latency is essential for maintaining the real-time 86.0 ms end-to-end processing speed of the tracking pipeline. Meanwhile, the downstream Re-ID clustering module demonstrates sufficient robustness to tolerate this small fraction of misleading frames without suffering from severe association failures.

5. CONCLUSIONS

In this study, we addressed a critical bottleneck in multi-object tracking, where downstream task performance is degraded by redundant and low-quality data. We first optimized the YOLO11m-CSA pedestrian detector to extract high-fidelity initial bounding boxes. Building on this robust perception baseline, we introduced a Dynamic Four-Stage Keyframe Extraction strategy that acts as a tracker-aware data purifier. By analyzing internal tracking states, including detection confidence and information entropy, it intelligently selects a sparse yet high-value set of keyframes.

Our experiments on MOTChallenge benchmarks yielded two crucial findings. First, the strategy achieves a data reduction rate of over 96.8% while dramatically enhancing tracking accuracy in extremely crowded scenes. On the challenging MOT20-02 sequence, the MOTA score improved from a near-failure baseline of 9.6% to 73.5%. This improvement stems from addressing the "garbage in, garbage out" problem by filtering pollutant frames that would otherwise corrupt feature vectors for the downstream Re-ID model.

This work demonstrates a highly adaptable algorithmic extension focused on intelligently refining the tracker output rather than modifying the underlying tracking mechanism. For future work, we identify two primary directions. First, we will improve the Tracklet

Integrity Assurance for very brief tracks. Second, we will address the system's residual vulnerability to complex background noise that mathematically bypasses our temporal and entropy filters. To resolve this, we will explore sophisticated quality metrics such as No-Reference Image Quality Assessment (NR-IQA) models and semantic-level filtering. This approach will overcome the inherent limitations of using Shannon entropy as a universal proxy for Re-ID sample quality, ensuring that statistical randomness aligns more accurately with actual perceptual quality.

REFERENCES

1. M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, Deep Learning for Person Re-Identification: A Survey and Outlook, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, Jun. 2022, doi: 10.1109/TPAMI.2021.3054775.
2. L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking," Apr. 08, 2015, *arXiv*: arXiv:1504.01942. doi: 10.48550/arXiv.1504.01942.
3. N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "BoT-SORT: Robust Associations Multi-Pedestrian Tracking," Jul. 07, 2022, *arXiv*: arXiv:2206.14651. doi: 10.48550/arXiv.2206.14651.
4. Y. Zhang *et al.*, "ByteTrack: Multi-Object Tracking by Associating Every Detection Box," Apr. 07, 2022, *arXiv*: arXiv:2110.06864. doi: 10.48550/arXiv.2110.06864.
5. H. Liang *et al.*, "KeyVideoLLM: Towards Large-scale Video Keyframe Selection," Aug. 10, 2024, *arXiv*: arXiv:2407.03104. doi: 10.48550/arXiv.2407.03104.
6. I. Erregue, K. Nasrollahi, and S. Escalera, "YOLO11-JDE: Fast and Accurate Multi-Object Tracking with Self-Supervised Re-ID," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, Feb. 2025, pp. 776–785. doi: 10.1109/WACVW65960.2025.00092.
7. J. Cai *et al.*, "MeMOT: Multi-Object Tracking with Memory," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 8080–8090. doi: 10.1109/CVPR52688.2022.00792.
8. S. Bharathi, M. Senthilarasi, and K. Hari, "Key frame extraction based on real-time person availability using YOLO," *J. Wirel. Mob. Netw. Ubiquitous Comput. Dependable Appl.*, vol. 14, no. 2, pp. 31–40, 2023.
9. A. Singh, L. Sanayai Meetei, S. M. Singh, T. D. Singh, and S. Bandyopadhyay, "An Efficient Keyframes Selection Based Framework for Video Captioning," in *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, S. Bandyopadhyay, S. L. Devi, and P. Bhattacharyya, Eds., National Institute of Technology Silchar, Silchar, India: NLP Association of India (NLPAI), Dec. 2021, pp. 240–250. Accessed: Jul. 04, 2025. [Online]. Available: https://aclanthology.org/2021.icon-main.29/
10. R. Roberts, J. P. Lewis, K. Anjyo, J. Seo, and Y. Seol, "Optimal and interactive keyframe selection for motion capture," *Comput. Vis. Media*, vol. 5, no. 2, pp. 171–191, Jun. 2019, doi: 10.1007/s41095-019-0138-z.

11. K. Nalaie, R. Xu, and R. Zheng, "DeepScale: Online Frame Size Adaptation for Multi-object Tracking on Smart Cameras and Edge Servers," in *2022 IEEE/ACM Seventh International Conference on Internet-of-Things Design and Implementation (IoTDI)*, May 2022, pp. 67–79. doi: 10.1109/IoTDI54339.2022.00010.
12. K. Nalaie and R. Zheng, "AttTrack: Online Deep Attention Transfer for Multi-object Tracking," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2023, pp. 1654–1663. doi: 10.1109/WACV56688.2023.00170.
13. X. Li, C. Chen, Y. Lou, M. Abdallah, K. T. Kim, and S. Bagchi, "HopTrack: A Real-time Multi-Object Tracking System for Embedded Devices," Nov. 01, 2024, *arXiv*: arXiv:2411.00608. doi: 10.48550/arXiv.2411.00608.
14. S. Zhang, J. Yang, J. Yin, Z. Luo, and J. Luan, "Q-Frame: Query-aware Frame Selection and Multi-Resolution Adaptation for Video-LLMs," Jun. 27, 2025, *arXiv*: arXiv:2506.22139. doi: 10.48550/arXiv.2506.22139.
15. J. Yoon and M.-K. Choi, "Exploring Video Frame Redundancies for Efficient Data Sampling and Annotation in Instance Segmentation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2023, pp. 3308–3317. doi: 10.1109/CVPRW59228.2023.00333.
16. W. Li, C. Chen, and K. Huang, "Absorb and Repel: Pseudo-Label Refinement for Intra-Camera Supervised Person Re-Identification," *IEEE Trans. Artif. Intell.*, vol. 5, no. 6, pp. 2884–2896, Jun. 2024, doi: 10.1109/TAI.2023.3327671.
17. Q. Hou, D. Zhou, and J. Feng, "Coordinate Attention for Efficient Mobile Network Design," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA: IEEE, Jun. 2021, pp. 13708–13717. doi: 10.1109/CVPR46437.2021.01350.
18. Glenn, J. YOLO11. Git Code. 2024. Available online: https://github.com/ultralytics/ultralytics (accessed on 10 October 2024).
19. M. Adžemović, "Deep Learning-Based Multi-Object Tracking: A Comprehensive Survey from Foundations to State-of-the-Art," Jun. 16, 2025, *arXiv*: arXiv:2506.13457. doi: 10.48550/arXiv.2506.13457.
20. A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*, Sep. 2016, pp. 3464–3468. doi: 10.1109/ICIP.2016.7533003.
21. Y. Pritch, A. Rav-Acha, and S. Peleg, "Nonchronological Video Synopsis and Indexing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1971–1984, Nov. 2008, doi: 10.1109/TPAMI.2008.29.
22. K. Nalaie, "Accelerating Multi-target Visual Tracking on Smart Edge Devices," Thesis, 2023. Accessed: Jul. 04, 2025. [Online]. Available: https://macsphere.mcmaster.ca/handle/11375/28771
23. X. Chu, A. Zheng, X. Zhang, and J. Sun, "Detection in Crowded Scenes: One Proposal, Multiple Predictions," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 12211–12220. doi: 10.1109/CVPR42600.2020.01223.
24. A. Vaswani *et al.*, "Attention Is All You Need," Aug. 02, 2023, *arXiv*: arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.

25. S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Sep. 2018.
26. Z. Wang *et al.*, "PC-YOLO11s: A Lightweight and Effective Feature Extraction Method for Small Target Image Detection," *Sensors*, vol. 25, no. 2, Art. no. 2, Jan. 2025, doi: 10.3390/s25020348.
27. S. Shao *et al.*, "CrowdHuman: A Benchmark for Detecting Human in a Crowd," *ArXiv*, Apr. 2018, Accessed: Mar. 17, 2025. [Online]. Available: https://www.semanticscholar.org/paper/CrowdHuman%3A-A-Benchmark-for-Detecting-Human-in-a-Shao-Zhao/03a65d274dc6caea94f6ab344e0b4969575327e3
28. S. Zhang, Y. Xie, J. Wan, H. Xia, S. Z. Li, and G. Guo, "WiderPerson: A Diverse Dataset for Dense Pedestrian Detection in the Wild," Sep. 25, 2019, *arXiv*: arXiv:1909.12118. doi: 10.48550/arXiv.1909.12118.
29. C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948, doi: 10.1002/j.1538-7305.1948.tb01338.x.
30. A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "MOT16: A Benchmark for Multi-Object Tracking," May 03, 2016, *arXiv*: arXiv:1603.00831. doi: 10.48550/arXiv.1603.00831.
31. F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera People Tracking with a Probabilistic Occupancy Map," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 267–282, Feb. 2008, doi: 10.1109/TPAMI.2007.1174.
32. L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable Person Re-identification: A Benchmark," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 1116–1124. doi: 10.1109/ICCV.2015.133.
33. E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance Measures and a Data Set for Multi-target, Multi-camera Tracking," G. Hua and H. Jégou, Eds., in Lecture Notes in Computer Science, vol. 9914. Cham: Springer International Publishing, 2016, pp. 17–35. doi: 10.1007/978-3-319-48881-3_2.
34. X. Han *et al.*, "Rethinking Sampling Strategies for Unsupervised Person Re-Identification," *IEEE Trans. Image Process.*, vol. 32, pp. 29–42, 2023, doi: 10.1109/TIP.2022.3224325.
35. G. Jocher, *YOLOv5 by Ultralytics*. (May 2020). doi: 10.5281/zenodo.3908559.
36. G. Jocher, J. Qiu, and A. Chaurasia, *Ultralytics YOLO*. (Jan. 2023). [Online]. Available: https://github.com/ultralytics/ultralytics
37. C.-Y. Wang, I.-H. Yeh, and H.-Y. Mark Liao, "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., Cham: Springer Nature Switzerland, 2025, pp. 1–21. doi: 10.1007/978-3-031-72751-1_1.
38. A. Wang *et al.*, "YOLOv10: Real-Time End-to-End Object Detection," 2024, doi: 10.48550/ARXIV.2405.14458.



Feng-Cheng Lin (林峰正) received his Ph.D. degree in Computer Science and Information Engineering from National Chi Nan University, Taiwan, in 2008. Currently, he is an associate professor in the Department of Information Engineering and Computer Science of Feng Chia University, Taichung, Taiwan. His research interests include cloud computing, deep learning, business intelligence, and multimedia applications.



Huan-Liang Elvin Lee (**李煥良**) was born in Kuching, Sarawak, Malaysia, in 1999. He received the B.S. and M.S. degrees in information engineering and computer science from Feng Chia University, Taichung, Taiwan, in 2022 and 2025, respectively. His research interests include computer vision, embedded AI, and software engineering.