

MixSORT: 적응형 복합 실시간 객체 추적 시스템

MixSORT: Adaptive Hybrid Real-time Tracking System

김 유 진¹, 강 경 태², 권 보 경^{1,*}
(Yujin Kim¹, Kyungtae Kang², and Bokyung Amy Kwon^{2,*})

¹Department of Applied Artificial Intelligence, Major in Bio Artificial Intelligence, Hanyang University

²Department of Artificial Intelligence, Hanyang University

Abstract: One of the challenges in multi-object tracking (MOT) is maintaining ID consistency under real-time constraints. In particular, existing trackers need to fuse IoU and appearance embeddings without incurring latency or ID switches in complex environments such as crowded scenes, long occlusions, and reappearances. To mitigate this, we propose MixSORT, a system that integrates the real-time efficiency of simple online and real-time tracking (SORT) with the ID stability of DeepSORT by actively adopting distance correlations for matching sequences. MixSORT consists of a score-aware, two-stage association pipeline in which high-confidence detections are matched by IoU, and the remaining candidates are refined using distance-correlation-based appearance matching. In experiments on the MOT17-11 validation sequence, MixSORT enhances ID stability while improving MOTP, IDF1, IDSW, and long-term tracking metrics compared to DeepSORT. Future research should consider developing approximate distance-correlation computations when the number of target identities increases due to computational intensity.

Keywords: multi-object tracking, SORT, DeepSORT, real-time optimization, active search, deep learning

I. INTRODUCTION

Multiple-Object Tracking (MOT) is a technology that consistently maintains the location and identity (ID) of multiple objects across frames in a video stream. It is essential for various applications such as autonomous driving, intelligent surveillance, and robot navigation [1,2]. The ideal goal of an MOT system is to simultaneously ensure real-time performance and ID consistency. However, due to complex factors in real-world environments like occlusion, crowd density, viewpoint changes, and dynamic backgrounds, frequent ID switches occur. Addressing these limitations requires sophistication in designing association rules, gating strategies, and managing the object lifecycle, including track creation, confirmation, and termination. Additionally, it has been reported that conducting feature-quality enhancement—via preprocessing and representation learning—in parallel can indirectly suppress ID switches, thereby improving appearance stability [3].

The history of object tracking research shows that Single Object Tracking (SOT) has long been a central focus of the field. SOT concentrates on accurately following a single target object and is traditionally categorized into generative and discriminative model-based approaches. Correlation filter-based methods laid the foundation for low-latency tracking by enabling lightweight online adaptation [4].

Subsequently, the combination of convolutional features and Efficient Convolution Operators (ECO) expanded robustness to background clutter and scale variations [5]. Siamese network-based methods provided strong appearance matching through correlation operations between a template and a search image, ensuring both simplicity and performance with end-to-end learning [6]. Transformer-based tracking methods effectively handled occlusion and speed changes by modeling long- and short-term dependencies with attention mechanisms [7]. These advancements have been extended to multi-object environments, where design principles such as template quality management, online gallery updates, and attention-based feature fusion contribute to enhancing inter-object discriminability and re-association capabilities. Recent MOT trend analyses comprehensively summarize these connections between SOT and MOT, as well as the resulting design space of online MOT systems [1,8-9].

Modern MOT is dominated by the tracking-by-detection paradigm as the standard approach. In this framework, a detector provides object candidate boxes for each frame, and the tracker then assigns these candidates to tracks and manages their lifecycle using dynamic motion prediction and global assignment. The Simple Online and Realtime Tracking (SORT) algorithm [10] achieved real-time performance with

* Corresponding Author

Manuscript received November 11, 2025; revised December 1, 2025; accepted December 18, 2025

김유진: 한양대학교 인공지능융합학과 대학원생(yujink@hanyang.ac.kr, ORCID[®] 0009-0007-6717-619X)

강경태: 한양대학교 인공지능학과 교수(ktkang@hanyang.ac.kr, ORCID[®] 0000-0002-6587-7044)

권보경: 한양대학교 인공지능학과 부교수(amykwon@hanyang.ac.kr, ORCID[®] 0000-0003-1632-0677)

※ 본 연구는 한국 연구 재단 (NRF) 기초 개인 연구 사업 및 정보 통신 기획 평가원 (IITP)의 지원에 의하여 연구되었음.

※ This research was supported by Basic Science Research Program of the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2023-00237241). This work was also partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2022-00155885, Artificial Intelligence Convergence Innovation Human Resources Development (Hanyang University ERICA)).

a simple structure combining Kalman prediction, Intersection over Union (IoU) based cost, and Hungarian assignment. However, due to the lack of appearance information, it was reported to have limitations in ID consistency in complex scenes. DeepSORT [11] improved upon this by introducing deep appearance embeddings, statistical gating, and a matching cascade, significantly enhancing ID consistency. However, this came at the cost of increased latency due to feature extraction and gallery management. Subsequent research has delicately balanced lightness and robustness through advancements such as camera motion compensation (BoT-SORT) [12], efficient utilization of low-confidence detections (ByteTrack) [13], appearance enhancement and optimized gallery operation (StrongSORT) [14], and observation-centric scheduling (OC-SORT) [15], while more recent SORT-style and motion-centric trackers such as Hybrid-SORT [16], UCMCTrack [17], WM-SORT [18], and ESC-SORT [19] continue to refine these trade-offs between geometric modeling, camera motion compensation, and association robustness.

Because the tracking-by-detection paradigm fundamentally hinges on detector quality, complementary work has also explored training high-performance detectors such as Faster R-CNN and YOLO families for real-time deployment [20,21]. In parallel, deep learning-based object detection has been actively applied to intelligent transport systems, enabling lane-level congestion estimation, anomalous event detection, and rule-violation monitoring from monocular traffic cameras [2]. Performance evaluations are conducted using metrics like MOTA, MOTP, IDF1, IDSW, and FP/FN, with standard benchmarks such as MOT17 and MOT20 employed to verify real-world applicability by including various challenges [22].

Although there is a large body of work on the tracking-by-detection paradigm, few algorithms directly address the consistency problem in long-term tracking. This paper proposes MixSORT to strengthen ID consistency while mitigating the trade-off between speed and accuracy. The core of the proposed method is a hierarchical pipeline that divides association into two stages based on detection confidence. In stage 1, IoU based global assignment is applied to high-confidence detections to suppress latency. In stage 2, for the remaining tracks and low-confidence detections, spatial incompatibilities are filtered using the mean and covariance of Kalman-filter predictions, followed by global optimization using distance correlation as an appearance similarity metric. Distance correlation is a statistical similarity measure that captures both linear and nonlinear dependencies, enabling selective and robust association even during periods of large appearance variation such as occlusion or re-emergence [23-25]. MixSORT is designed to simultaneously achieve continuity maintenance and identity stabilization under online conditions by saving time in stage 1 and performing refined comparisons only for necessary pairs in stage 2.

II. RELATED STUDIES

1. Tracking-by-Detection Overview

Currently, MOT largely follows a detection-based procedure that takes detection results as input to update tracks. In each frame, when the detector outputs bounding boxes and confidence scores, the tracker performs data association between consecutive frames, followed by state estimation and track management. The advantage of this approach lies in its modularity. It offers benefits in terms of reproducibility and scalability because it enables the independent improvement of individual modules such as detection, association, state

estimation, and track management [10,11].

The characteristics of the detection stage, such as the rates of false positives and false negatives, and the distribution of low-confidence detections, are reflected in the association design. Specifically, these become core factors that define the cost function design, gating intensity, and track management scheduling strategies. Detectors have evolved into methods such as anchor-based Faster R-CNN [20], single-stage YOLO detectors [21], and transformer-based DETR and Deformable-DETR [26,27]. Post-processing techniques like non-maximum suppression (NMS), soft-NMS, and Optimal Transport Assignment (OTA), as well as Distance Intersection over Union (DioU) and Complete Intersection over Union (CioU) family loss functions and feature parameter design, significantly impact the final precision and recall. Particularly, in adverse conditions such as fog, night, or backlighting, low-confidence detections increase, leading to a rapid expansion of the association search space and a higher risk of mismatches. Therefore, techniques such as dehazing, image restoration, domain adaptation, and test time adaptation (TTA) are used concurrently. These pre-processing and adaptation methods enhance the possibility of recovering true positives in low-confidence regions and provide a foundation for subsequent score-aware cascade association to function effectively [13].

2. Association and Online Trackers

The success of online MOT systems is contingent upon the design of data association strategies. Given real-time constraints, most methodologies perform global one-to-one assignments by integrating IoU based proximity via Kalman filters with detection confidence derived from appearance descriptors for target identity. SORT [10] achieved simple and fast matching through a Hungarian algorithm assignment method based on the minimum cost of IoU predictions via Kalman filters; however, it reported limitations, namely an increase in identity switches during close proximity interactions or long-term occlusions. As a complementary approach, DeepSORT [11] enhanced identity stability by incorporating an appearance descriptor and adjusting matching priorities to improve the accuracy of re-identification for target identities. Subsequent research has proceeded in a direction that compromises and expands upon the advantages of these two methods. BoT-SORT [12] restored the reliability of IoU costs by alleviating inter-frame geometric variations from moving cameras through global motion compensation (GMC). ByteTrack [13] structurally absorbed true positives in low-confidence regions while suppressing mis-associations using a two-stage score-aware strategy of high-confidence first matching followed by low-confidence matching. StrongSORT [14] reinforced the temporal consistency of appearance costs via a robust re-identification backbone, gallery quality gating, and budget management. OC-SORT [15] compensated for the jitter and non-linear motion vulnerabilities of pure IoU-based estimation by utilizing observation-centric updates, as well as velocity and direction information. These trackers commonly reduce error propagation and computational load by pre-masking impossible combinations through gating, and maintain a balance of accuracy, robustness, and cost in online environments by combining motion, appearance, and detection score information in a stepwise or weighted manner.

In scenes with small bounding boxes at long range, the IoU between consecutive frames tends toward zero, and many online trackers adopt nearest-neighbor association based on the Euclidean distance between bounding-box centers as a geometric fallback [28]. This center-distance heuristic relies exclusively on two-dimensional coordinates and has

several limitations. A one-pixel shift in the image plane can correspond to a large change in physical space, so camera motion, perspective distortion, and localization noise strongly affect the measured distance. In crowded scenes where many objects move along similar trajectories, multiple targets produce similar center distances and identity cues disappear. Appearance information accumulated along each track over time also remains unused at the matching stage.

The distance correlation (dCor) based association cost used in this study addresses these weaknesses of geometry-only matching. This method compares the high-dimensional embedding distribution of a candidate detection through pairwise distance statistics and measures nonlinear dependence that remains stable under monotone rescaling [23]. Re-identification embeddings capture clothing color, silhouette, and local texture, and dCor uses these cues to preserve reliable identities for small and distant objects. In such regimes, geometric proximity causes center-distance matching to merge neighboring trajectories, whereas dCor-based appearance association provides additional discriminative signals. This analysis motivates a design that combines IoU-based geometric information with dCor-based appearance cost, and Section IV describes the resulting structure and algorithms in detail.

3. Real-Time Optimization and Appearance Representation

The tracking quality of a real-time system is largely determined by the track lifecycle policy and inference light-weighting. Following track initialization, subsequent stages are typically classified into a *'tentative'* state (where the matching status is not yet confirmed), a *'confirmed'* state (due to a certain period of consistent matching), and a *'missed'* state (where the target identity has not been found for a certain duration). Hyperparameters such as appearance distance thresholds, along with confidence thresholds, as well as measurements, such as IoU determine the balance of tracking quality due to trade-offs concerning identification stability, processing latency, and the rates of false positives and false negatives for the target identity.

Re-identification typically involves a procedure of comparing distances using L2-normalized embeddings derived from CNNs or Vision Transformers (ViTs). In the training phase, multi-loss configurations that combine triplet loss, contrastive loss, and softmax classification loss, along with hard sample mining, label refinement, and domain adaptation, are known to contribute to generalization performance.

Furthermore, during the inference phase, system optimizations are jointly applied to reduce computational load. These optimizations include extracting embeddings primarily for candidate pairs that pass through a gate, prototype aggregation based on recent or representative values, preventing memory inflow for costs below a certain threshold, and employing techniques such as batching, caching, mixed precision, quantization, and kernel fusion. The combination of this score-aware cascade and gating-based masking presents a practical solution that simultaneously pursues the robust identification characteristic of the DeepSORT family, and the low-confidence recall advantages of the ByteTrack family.

III. Brief Description: SORT vs. DeepSORT

In multi-object tracking, both SORT and DeepSORT share the same fundamental architecture: they utilize a Kalman filter to predict track states and employ the Hungarian algorithm for global one-to-one assignment between predicted tracks and current frame detections. However, the core difference between the two algorithms is starkly

evident in their data association costs and gating strategies. Fig. 1 visually summarizes these differences.

SORT employs a single-stage, IoU-based matching approach. It is configured as a highly streamlined pipeline that calculates association costs solely using the intersection ratio between predicted and detected bounding boxes, followed by assignment via the Hungarian algorithm. It applies only to an IoU mask, which excludes pairs with an IoU smaller than a threshold (τ_{IoU}) from the cost matrix, without using additional gating based on statistical uncertainty. This structure offers advantages such as low computational load and excellent real-time performance. DeepSORT enhances the representational power of associations by combining appearance-based signals and statistical gating within SORT's basic framework. It extracts appearance embeddings from detection regions, maintains them in a per-track gallery, and uses appearance similarity to calculate the association cost for matching. Simultaneously, it uses Mahalanobis distance gating, derived from the mean and covariance of Kalman predictions, to preemptively constrain physically impossible or inappropriate candidates. The final assignment proceeds via a cascaded matching procedure that prioritizes recently updated tracks. While SORT suffers from an increased number of ID switches in scenarios like crowd intersections, reappearance after prolonged occlusion, or camera motion, DeepSORT significantly improves ID consistency as well as IDFI in the same scenarios due to this cascade matching process with an enhanced association descriptor compared to SORT. However, the latency in DeepSORT increases due to the cost of embedding extraction and gallery management, and additional training is required if the domain suitability of the embeddings is low.

Practical application dictates that SORT is a reasonable choice when the scene is simple, detection quality is high, and the latency budget is very strict. Conversely, DeepSORT's appearance augmentation provides an advantage in situations with frequent mutual occlusion and intersection, or when geometric information is unstable due to camera motion. Recently, a hybrid pipeline that compromises between the two methods has become widely used: it rapidly absorbs high-confidence detections using IoU to suppress latency and then applies precise correction using appearance-based costs for the remaining items in a second stage. Ultimately, the two frameworks are choices within the same global optimization structure, but they place the balance point between the expressiveness of the association signal and processing latency at different positions. It is desirable to select a single method or design a hybrid configuration depending on the application's latency limits, ID preservation requirements, and scene characteristics.

IV. HYBRID REAL-TIME OBJECT TRACKING SYSTEM

In this paper, we propose MixSORT, an adaptive hybrid real-time object tracking system. MixSORT is based on the aforementioned SORT [10], partially incorporates the appearance identification process of DeepSORT [11], and utilizes a distance correlation metric aiming to increase the consistency of ID especially for a long tracking sequence by enhancing tracking selectivity.

Online tracking systems generally assume uncalibrated cameras and the absence of ego-motion information, and MixSORT also operates on a frame-by-frame basis within this environment. As shown in Fig. 1, MixSORT maintains a framework of IoU distance in SORT, and an additional appearance descriptor in DeepSORT for the same purpose as these two tracking systems. When a frame is ingested by the system,

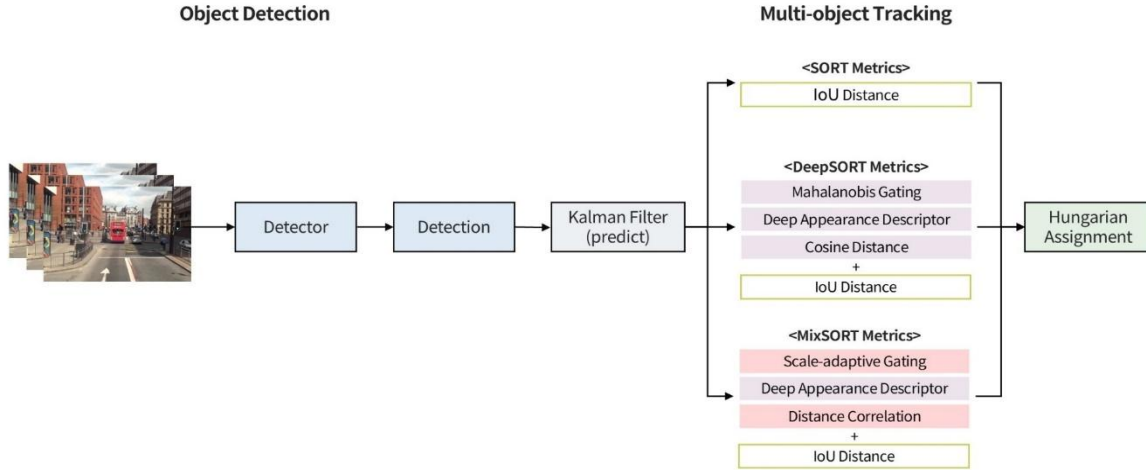


그림 1. 검출 기반 온라인 다중 객체 추적 파이프라인. 좌측은 객체 검출 단계를, 우측은 칼만 필터 예측과 SORT, DeepSORT 및 제안하는 MixSORT의 연관 지표 구성을 보여주며, MixSORT의 스케일 적응 게이팅과 거리 상관 기반 외관 비용 추가를 강조한다.

Fig. 1. Detection-based online multi-object tracking pipeline. The left panel shows the object detection stage, while the right panel illustrates Kalman prediction and the association metrics of SORT, DeepSORT, and the proposed MixSORT, highlighting MixSORT's scale-adaptive gating and distance-correlation-based appearance cost.

Kalman filter prediction is performed for all active tracks. However, MixSORT has a two-stage matching structure to enhance the ID selectivity based on a confidence level (based on IoU score). When the confidence level of a track is defined as highly confident, the matching is solely based on IoU distance, but the matching is based on distance correlation of the appearance descriptor considering an elliptical gating mechanism otherwise, where the distance correlation measures a probabilistic association between two vectors considering either linear or non-linear association independent of scale [23]. Subsequently, track states are updated and managed according to the matching results.

1. State Estimation

We continuously estimate the object state by applying a Kalman filter model based on the latest position and size information of the tracked bounding boxes [29,30]. While aspect ratio or area could be used as state variables, as in previous studies [10,11], this approach has limitations in directly predicting and estimating the fine-grained changes of the actual bounding box coordinates, even if advantageous for estimating macroscopic changes such as the shape or size of the bounding box. Therefore, in this study, we represent the motion state using a linear constant velocity model, utilizing the bounding box's center coordinates (x, y) , width w , and height h . The state vector is as follows.

$$X = [x, y, w, h, \dot{x}, \dot{y}, \dot{w}, \dot{h}] \quad (1)$$

The initial state is set by the first observation, with velocity components initialized to zero. The state error covariance, P and the process and observation noises, Q and R are constructed as diagonal components using axis-specific standard deviation coefficients proportional to the size of the bounding box and are recalculated every frame based on the current width and height based on constant velocity assumption. When an observation is received, the standard Kalman filter update procedure is followed; if no observation is available, only the prediction step is performed. For numerical stability, a minimum lower bound is applied to the width and height. If a bounding box is expected to collapse in the next frame, the corresponding velocity

components are clipped to zero. During initialization, the variances of the size and velocity axes are set relatively large to prevent oversharing (overconfidence), and whenever an observation is received, the observation noise covariance is recalculated based on that observation's width and height.

2. Appearance Descriptor

Appearance feature extraction follows the deep appearance embedding procedure proposed in DeepSORT, where a CNN is pre-trained offline using the large-scale dataset MARS (approximately 1.1 million images, 11,261 identities) for pedestrian re-identification [31]. The network follows a Wide Residual Network [32] backbone consisting of two initial convolutional layers and six residual blocks, yielding a 128-dimensional feature embedding in the upper layers.

Batch normalization and L2 normalization are sequentially applied to the resulting embeddings to project them to unit norm, thereby securing appearance representations that are less sensitive to scale variations [33]. In the online stage, detection boxes are normalized to a fixed resolution and then processed through the same preprocessing steps for inference. This process generates a normalized 128-dimensional appearance feature vector every frame.

3. Distance Correlation

In this study, we adapt a distance correlation as an appearance similarity metric. Distance correlation, $R(X, Y)$ is a probabilistic measure representing either linear or a non-linear association between two vectors, which can be calculated as (2). $R^2(X, Y)$ becomes zero only when the two random vectors are independent, and its value ranges from $[0, 1]$, representing the strength of dependency. The definition of distance correlation is as follows.

$$R^2(X, Y) = \begin{cases} \frac{v^2(X, Y)}{\sqrt{v^2(X)v^2(Y)}}, & v^2(X)v^2(Y) > 0 \\ 0, & v^2(X)v^2(Y) = 0 \end{cases} \quad (2)$$

Here, $v^2(X, Y)$ is the distance covariance, which is defined as follows using the joint characteristic function, $f_{XY}(t, s)$ and the marginal characteristic functions $f_X(t)$ and $f_Y(s)$.

$$v^2(X, Y) = \|f_{xy}(t, s) - f_x(t)f_y(s)\|^2 \quad (3)$$

In the special case of a bivariate normal distribution, the relationship with the traditional correlation coefficient ρ is such that $R = 1$ when $|\rho| = 1$, and the following relationship is satisfied.

$$|\rho(X, Y)| \leq R(X, Y) \leq 1 \quad (4)$$

Therefore, distance correlation can be considered a generalized concept of the Pearson correlation coefficient, capturing not only linear relationships but also nonlinear dependencies [23].

The actual calculation uses the sample distance correlation of the double-centered Euclidean distance matrices. We compute the pairwise Euclidean distance matrices for the set of detection embeddings and the set of track gallery embeddings, obtain the sample distance covariance and sample distance variance through double-centering (removing row and column means), and then acquire the value $R \in [0, 1]$ via the normalization in (2). The sample distance correlation operates stably even with high-dimensional embeddings, and the measured latency can be sufficiently suppressed by combining efficient computation techniques such as mini-batch calculation, partial sampling, cumulative updating, and approximation methods [25].

In this study, we define the appearance-based cost as 1-R, utilizing the characteristic that dependency is stronger when R is larger. We also implement an adaptive association strategy that actively selects and combines appearance-based metrics and IoU-based metrics depending on the situation of the tracklet.

4. Gating Parameters

To pre-emptively eliminate candidates that are physically unlikely to be valid between the object's predicted state and the newly received measurement, we use scale-adaptive gating based on the elliptical Euclidean principle. Gating is the procedure of retaining only those candidates for computation whose distance measurement between the track and the detection candidate pair is below a certain threshold. An elliptical gate is used because circular or fixed-scale Euclidean gates can cause proportional distortion when object size changes or aspect ratios differ. Specifically, an elliptical gate is positioned relative to each track's predicted box center (x_p, y_p) , with its minor and major axes proportional to the predicted width (w), and height (h). The normalized squared distance to the detection center (x_d, y_d) is defined as follows:

$$d^2 = \left(\frac{x_d - x_p}{\alpha w}\right)^2 + \left(\frac{y_d - y_p}{\beta h}\right)^2 \quad (5)$$

Here, α and β are scale adjustment coefficient for width and height, respectively where $\alpha, \beta > 0$. In Eq (5), a detection is maintained as an association candidate only if $d^2 \leq d_{th}$. This implements anisotropic (elliptical) gating that is stricter for smaller targets and more lenient for larger targets, even with the same pixel error. The size of the ellipse is updated every frame with the predicted w and h to immediately reflect size changes due to perspective changes, zoom, and acceleration. In implementation, track-detection pairs where $d^2 > d_{th}$ are masked in the cost matrix using a large constant value, where d_{th} is adjusted by the validation procedure. For those entries which pass this gate, the appearance costs are computed and then passed to the Hungarian optimization algorithm. This scale-adaptive elliptical gating effectively reduces false associations and lowers the noise in the association cost matrix in scenes where objects of different sizes and

shapes coexist or where object size changes significantly due to perspective effects, thereby reducing the computational burden of subsequent optimization and appearance cost calculations. While α and β basically use the same value, they can be set differently for each axis depending on scene characteristics, such as vertical and horizontal uncertainty deviations.

5. Track Handling

To manage track confidence, MixSORT uses three finite states: *tentative*, *confirmed*, and *deleted*. When a frame is input to the system, a Kalman filter prediction is first performed for all active tracks, and a global matching process performs one-to-one assignment with the detection results. Unmatched detections are immediately initialized as tentative tracks based on the observation. At this point, the mean of the initial state is set to the first observation, and the velocity components are set to zero. Each track is managed with two counters: *hits*, which represents the number of successful matches so far, and *time_since_update*, which indicates the number of frames since the last successful association. Tentative tracks must pass a short confirmation period; if they meet *min_hits* (e.g., three consecutive successful matches), they are promoted to the confirmed state. If a tentative track fails to match even once during the initial confirmation period, it is immediately deleted.

For successfully matched tracks, the Kalman filter is updated, the number of hits is incremented, and *time_since_update* is reset to 0. The latest appearance embedding is added to the track gallery. Conversely, tracks that fail to match are only predicted without being updated, and *time_since_update* is incremented by 1. Regardless of the track's state, if *time_since_update* exceeds a predefined maximum allowed un-updated frames, *max_age*, the track is considered to have left the scene and is terminated. New tracks are not regenerated from detections that have already been associated in the same frame, and track identifiers are not reused throughout the sequence. This simple state machine and counter-based policy are advantageous for their easy implementation in real-world online environments and clear tuning points.

6. Assignment and Matching

Data association is separated into high-confidence (high) and low-confidence (low) based on detection confidence, and matching is performed sequentially through a two-stage global optimization. Both stages construct a cost matrix and find the optimal assignment that minimizes the total cost using the Hungarian algorithm [34].

In the first stage, the cost is defined based on the IoU between the predicted tracks and only the high-confidence detections.

$$C_{ij}^{IoU} = 1 - \text{IoU}(T_i, D_j) \quad (6)$$

Additionally, global matching is performed only for pairs that pass the IoU threshold mask.

$$M_{ij}^{IoU} = 1[\text{IoU}(T_i, D_j) \geq \tau_{IoU}] \quad (7)$$

Matched pairs are immediately confirmed with a state update, and the IoU-unmatched tracks, detections, and low-confidence detections are carried over to the second stage. This process aligns with a score-aware strategy that conservatively absorbs high-confidence detections first and then handles remaining and low-confidence detections in a subsequent stage [13].

The second stage handles the remaining unmatched tracks and detections from the first stage, along with low-confidence detections.

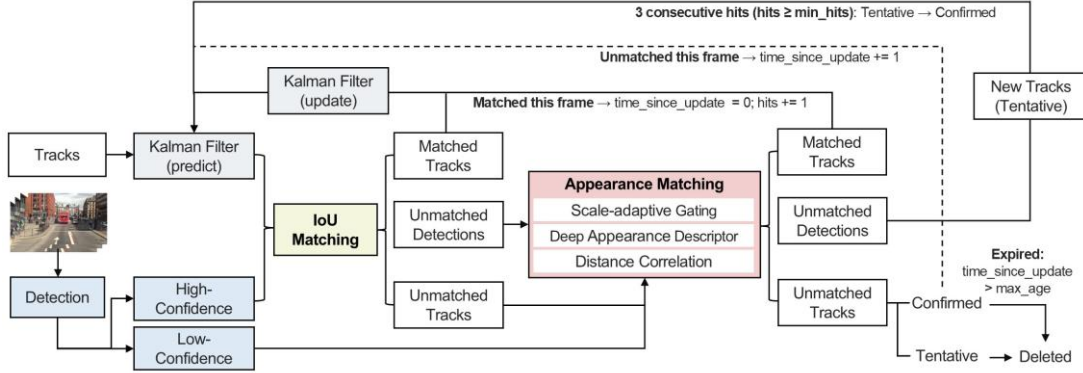


그림 2. 제안된 MixSORT 파이프라인 개요. 검출 신뢰도에 따라 IoU 기반 1단계 연관과 스케일 적응 게이팅, 딥 외관 기술자, 거리 상관을 이용한 2단계 외관 연관을 수행하고, tentative-confirmed-deleted 상태 전이를 통해 트랙을 관리한다.

Fig. 2. Overview of the proposed MixSORT pipeline. A confidence-aware two-stage association first matches high-confidence detections using IoU and then performs appearance matching with scale-adaptive gating, deep appearance descriptors, and distance correlation while managing tracks through tentative, confirmed, and deleted states.

This stage focuses on improving the precision of appearance information. First, elliptical Euclidean gating is applied to pre-emptively exclude spatially inappropriate combinations. For candidates that pass the gate, an appearance-based matching is performed using the distance correlation similarity between the track gallery embeddings and the detection embeddings as the cost. The distance correlation R between the track embedding set F_i and the detection embedding f_j is as follows.

$$R(F_i, f_j) = \text{dCor}(F_i, f_j) \quad (8)$$

Here, the appearance-based cost is defined as $1 - R$.

$$C_{ij}^{\text{app}} = 1 - R(F_i, f_j) \quad (9)$$

Additionally, by setting a quality threshold, a cost matrix is constructed using only the pairs that satisfy it, and global optimization is performed using the Hungarian algorithm.

$$M_{ij}^{\text{app}} = 1[C_{ij}^{\text{app}} \leq \tau_{\text{app}}] \quad (10)$$

By actively combining detection confidence, IoU, and appearance information and separating the process into two stages, this configuration allows MixSORT to minimize identity switches and secure re-detection recovery even in crowded scenes.

7. Architecture

At its core, MixSORT employs a score-aware, two-stage matching process for every frame. When a new frame is ingested, the active tracks from the previous frame are predicted by a Kalman filter and propagated to the next state. The detection results are branched into high-confidence and low-confidence groups according to their confidence scores. Fig. 2 illustrates the entire pipeline.

Data association is divided into two sequential stages in MixSORT. Stage 1 performs IoU-based matching between high-confidence detections and predicted tracks to minimize initial computational latency. Stage 2 performs precise matching for the remaining tracks and detections, including low-confidence detections, by first using spatial gating to eliminate implausible candidates, and then applying distance correlation-based appearance costs. This hierarchical structure ensures both real-time processing and ID consistency in scenes with crowding, occlusion, and viewpoint changes, by quickly assigning high-

confidence items and precisely correcting the remaining items with appearance information. Track management uses tentative, confirmed, and deleted states, and the latest embedding is added to the gallery upon update to increase the reliability of subsequent matching.

V. EXPERIMENTS

1. Dataset

The experiments in this study were conducted using the MOT17 dataset, a standard benchmark in the field of multi-object tracking. MOT17 consists of 14 video sequences filmed in various urban environments such as intersections, roads, and commercial districts, and it includes both static and moving camera scenarios. The dataset contains a wide range of challenges that make tracking difficult in real-world situations, such as changes in lighting and background, partial and full occlusion, differences in crowd density, and distant pedestrians. Each frame provides official annotations with bounding boxes and consistent object IDs, allowing for a fair performance comparison between algorithms [22,35].

Instead of directly using the official test split for final submission, a fixed validation sequence was separated and utilized from within the training split to stably conduct iterative ablation and hyperparameter search. Specifically, MOT17-02, MOT17-04, MOT17-05, MOT17-09, MOT17-10, and MOT17-13 were used as training sequences, while MOT17-11 was reserved as a separate validation sequence. This configuration allows for a consistent comparison of various settings under the same validation conditions without external server submission, and it naturally extends to the procedure of submitting the same model to the official test split if needed.

All experiments were conducted on the single pedestrian class, and the original resolution, frame rate, frame order, and encoding of each sequence were preserved. Only the official annotations provided by MOT17 were used for training and validation. An internal benchmark procedure was followed, where parameters were estimated using the training sequences, and then performance was compared and a model was selected using the separated MOT17-11 sequence. Unless otherwise noted, the results presented in the main text were derived using MOT17-11 as the validation set.

2. Evaluation Metrics

To objectively compare and analyze the performance of the MOT

system, this study utilized comprehensive evaluation metrics based on the CLEAR MOT framework [36]. The common evaluation metrics adopted were MOTA, MOTP, IDF1, IDSW, FP, FN, Frag, MT, and ML. These metrics correspond to the framework, allowing for the quantitative evaluation of the design goals of the MixSORT methodology: improving accuracy, stabilizing IDs, suppressing unnecessary detections and missed detections, and reducing track fragmentation. In the main text, we describe the core metrics among these—MOTA, MOTP, IDF1, IDSW, MT, and ML—while FP, FN, and Frag are also presented in the table to provide a complementary understanding of the system's error scale and track fragmentation patterns.

1) The Multiple Object Tracking Accuracy (MOTA):

$$\text{MOTA} = 1 - \frac{\sum_t (\text{FN}_t + \text{FP}_t + \text{IDSW}_t)}{\sum_t \text{GT}_t} \quad (11)$$

MOTA (Multi-Object Tracking Accuracy) is a representative metric that comprehensively indicates tracking performance. It is calculated by considering all missed detections (FN), false positives (FP), and ID switches (IDSW) at frame t . GT_t denotes the number of ground truth objects present in frame t . A value closer to 1 indicates higher overall tracking quality. Therefore, MOTA is influenced by both detection and association performance, and it shows how accurately the tracker identifies the number of objects and maintains their consistency.

2) The Multiple Object Tracking Precision (MOTP):

$$\text{MOTP} = \frac{\sum_{t,i} (1 - \text{IoU}_i^t)}{\sum_t c_t} \quad (12)$$

In this study, MOTP is used as a metric to represent the positional consistency of matched object pairs, where a higher value indicates better performance. IoU_i^t represents the IoU between the predicted bounding box and the ground truth bounding box of the i -th matched object in frame t , and c_t denotes the number of matches in that frame. Since this metric is calculated only for matched samples, it is not affected by the number of false negatives, false positives, or ID switches, and is used solely to compare the precision of position estimation. We based our work on the definition of CLEAR MOT [37], but re-expressed the py-motmetrics results, *Python* library, in an IoU format to present them in a consistent manner.

3) ID Switch (IDSW):

IDSW is defined as the number of times a single ground truth object is incorrectly assigned a different tracking ID during the tracking process. This metric directly demonstrates how stably the tracking algorithm maintains the object's identity, and a lower value indicates improved performance. IDSW primarily increases in scenarios involving frequent occlusion, close movement, and appearance changes; thus, it is used as a key metric when evaluating the effectiveness of appearance-based association modules.

4) ID F1-Score (IDF1):

$$\text{IDF1} = \frac{2 \cdot \text{IDP} \cdot \text{IDR}}{\text{IDP} + \text{IDR}} \quad (13)$$

IDF1 is a metric that integrates ID-oriented precision (IDP, ID Precision) and recall (IDR, ID Recall) using the harmonic mean to evaluate how long the same object is tracked with the same ID. While MOTA (Multiple Object Tracking Accuracy) primarily evaluates "how many objects were correctly detected," IDF1 assesses "how cleanly the IDs were maintained among the correctly detected objects." A higher

value for this metric indicates superior ID consistency.

5) Mostly Tracked (MT):

It represents the ratio of object tracks that were successfully tracked for more than 80% of their entire lifecycle. This metric directly demonstrates the tracker's ability to maintain an object stably over a long duration, in other words, the robustness of the track. By measuring the long-term tracking success rate, one can evaluate how well the tracking system handles complex scenarios.

6) Mostly Lost (ML):

It represents the ratio of object tracks that were successfully tracked for less than 20% of their entire lifecycle. This metric measures how many tracks the tracker observes only for a short time before failing. A high ML ratio indicates that the tracker lacks the recovery ability to re-detect objects after temporary occlusions, resulting in tracks being lost easily.

VI. RESULTS

1. Selection of Object Detectors Based on Performance Comparison

The final performance of multi-object tracking is directly dependent on the quality of the object detector, which is the initial input stage. In detection-based tracking, recall is crucial for creating new tracks and maintaining existing ones, while precision is a key indicator for suppressing false associations. In this study, to ensure system robustness, we conducted a comparison and analysis of DPM, Faster R-CNN (FRCNN), and YOLOv8 to select the final detector. A total of 9,436 GT boxes were used in the experiment, and data consistency was verified as each detector satisfied $\text{TP} + \text{FN} = 9,436$.

The quantitative results are summarized in Table 1. DPM performed the worst with 4,659 TPs and a recall of 0.4937. This means a very high FN of 4,777, likely causing track interruptions and frequent ID re-initializations during actual system application. FRCNN most effectively suppressed FPs with a precision of 0.9696, reducing the risk of incorrect associations during the linking stage. YOLOv8 had slightly higher FPs than FRCNN, making it less favorable in terms of precision, but its recall of 0.6141 resulted in the fewest missed objects among the three models. Consequently, YOLOv8 achieved an F1-score of 0.7398, showing a slight edge in the overall metric. In summary, we confirmed that YOLOv8 and FRCNN have contrasting characteristics: YOLOv8 focuses on recall (not missing detections), and FRCNN focuses on precision (not making false detections).

The same trend is observed in Fig. 3. YOLOv8 captures more pedestrians in long-distance scenarios and partial occlusion situations, reducing omissions, while FRCNN uses a conservative suppression strategy to aggressively remove low-confidence candidates, thus lowering FPs, but some TPs are missed. DPM frequently experiences object omissions in adverse conditions such as dense crowds, reflections, and motion blur.

Since this study aims to ensure track continuity (minimizing interruptions and re-initializations), YOLOv8, which reduces

표 1. MOT17-11에서의 객체 검출기 성능 비교.

Table 1. Comparison of object detector performance on MOT17-11.

Detector	precision	recall	F1-score	accuracy
DPM	0.9297	0.4937	0.6427	0.4736
FRCNN	0.9696	0.5955	0.7379	0.5846
YOLOv8	0.9300	0.6141	0.7397	0.5870

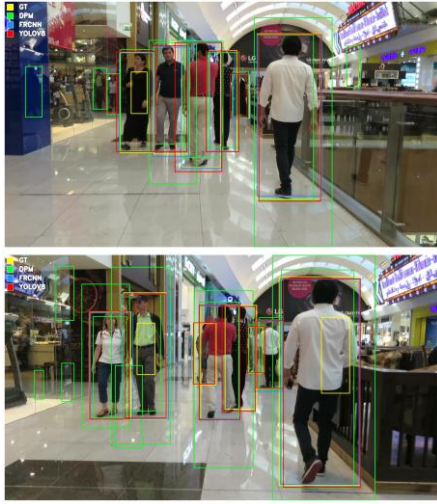


그림 3. 동일 프레임에서 GT와 DPM, FRCNN, YOLOv8의 보행자 검출 결과 비교.

Fig. 3. Comparison of pedestrian detection results of GT, DPM, FRCNN, and YOLOv8 in the same frame.

omissions with high recall, is the most suitable. YOLOv8 directly contributes to trajectory maintenance stability and showed superior performance in F1-score, confirming its advantage in overall balance.

2. Tracking Performance Comparison

In this study, we performed a comparative analysis of real-time tracking algorithms, including SORT [8], DeepSORT [9], and the proposed MixSORT, using standard evaluation metrics on the MOT-11 validation sequence. According to Table 2, the MOTA, which represents overall accuracy, is identical for DeepSORT and MixSORT at 60.1 each, which is 1.8 percentage points higher than SORT's 58.3. MixSORT had the highest MOTP (position precision metric) at 83.7, suggesting relatively superior bounding box location approximation in the matched segments. In terms of ID consistency, MixSORT showed the highest IDF1 at 59.2 and the lowest IDSW at 32, stably maintaining the ID of the same object. Long-term tracking metrics also favor MixSORT, with MT ratio of 30.7% and ML ratio of 44.0%, demonstrating excellent track maintenance and a low tendency for track loss. Regarding error composition, FN recorded the minimum value at 3506, but FP 224 and Frag 112 showed relatively high values. These results are interpreted as aggressive track activation and reactivation strategies for reducing omissions but partially accompanying increased over-detection and fragmentation. Nevertheless, the simultaneous improvement in IDF1, reduction in IDSW, and enhancement of MOTP and MT while maintaining the same MOTA level clearly demonstrate that the proposed association method and track management work towards prioritizing and strengthening ID stability and position precision.

Fig. 4 presents an enlarged view of a scene in which a target reappears after occlusion, enabling a detailed comparison of the results from each tracker. The figure visualizes the tracking outputs, where different box colors denote distinct tracking IDs, and red boxes indicate locations of ID switches. Unlike SORT and DeepSORT, which assigned new IDs to the same individual upon reappearance, the proposed MixSORT successfully maintained and restored the original ID. This observation demonstrates MixSORT's ability to ensure trajectory continuity across occlusion and reappearance, based on a two-stage association procedure, further supporting the previously reported improvements in IDF1 and reductions in IDSW.

On the other hand, the introduction of a distance-correlation based appearance association increases computational cost; when the number of candidate pairs is k , distance correlation requires pairwise computation of L2 distance at first, which results in $O(k^2)$ proportionally to the number of candidate pairs, potentially causing latency in applications with stringent real-time constraints. This study primarily focused on performance comparison; however, future work will explore strategies to alleviate such challenges -such as candidate reduction, approximate distance correlation computation, and lightweight batch processing- as well as post-processing techniques aimed at balancing different metrics, including conservative reactivation, fragmented track merging, and selective handling of low-confidence detections. Furthermore, since the present evaluation was conducted on a single validation sequence, it is necessary to extend the experiments to additional sequences encompassing diverse scene compositions, crowd densities, and occlusion patterns to assess the generalizability of the proposed approach.

VII. CONCLUSION

This paper proposes MixSORT, a novel online multi-object tracking framework designed to reconcile the persistent trade-off between ID consistency and real-time performance. MixSORT aims to integrate the real-time efficiency of SORT with the appearance-based ID stability of DeepSORT by introducing a two-stage matching pipeline guided by detection confidence. By adopting distance correlation as the appearance association cost, the method captures nonlinear dependencies in the embedding space, thereby enhancing ID recovery capability.

Experimental evaluations on the MOT17-11 validation sequence confirm the effectiveness of MixSORT compared to SORT and DeepSORT. MixSORT achieved a MOTA of 60.1%, comparable to DeepSORT, while improving MOTP and MT to 83.7% and 30.7%, respectively. In terms of ID consistency, it attained the highest IDF1 score of 59.2% and the lowest number of IDSW (32), demonstrating simultaneous improvements in both localization accuracy and trajectory continuity. Qualitative analysis further verified that MixSORT successfully preserved the same ID after occlusion and reappearance, consistent with the observed gains in IDF1 and reductions in IDSW. These results indicate that MixSORT enhances ID

표 2. MOT17-11 검증 시퀀스에서 온라인 추적기 성능 비교(↑: 높을수록 우수, ↓: 낮을수록 우수).

Table 2. Comparison of online trackers on the MOT17-11 validation sequence (↑: higher is better, ↓: lower is better).

Method		MOTA↑	MOTP↑	IDF1↑	IDSW↓	MT↑	ML↓	FP↓	FN↓	Frag↓
SORT [10]	ONLINE	58.3	83.4	54.4	67	24.0%	46.7%	215	3653	81
DeepSORT [11]	ONLINE	60.1	82.9	55.0	37	26.7%	45.3%	197	3527	91
MixSORT (Proposed)	ONLINE	60.1	83.7	59.2	32	30.7%	44.0%	224	3506	112



그림 4. 가림 후 재등장 장면의 ID 전환 비교. (a) SORT. (b) DeepSORT. (c) MixSORT. 제안하는 MixSORT는 동일 보행자의 ID를 유지하여, SORT와 DeepSORT에서 발생하는 ID 전환을 완화한다.

Fig. 4. ID switches in an occlusion-reappearance case. (a) SORT. (b) DeepSORT. (c) MixSORT. The proposed MixSORT preserves the identity of the same pedestrian, reducing ID switches that occur in SORT and DeepSORT.

stability and long-term trajectory maintenance while maintaining comparable overall accuracy (MOTA).

However, the computation of distance correlation and gallery access increases with the number of candidate pairs, potentially limiting real-time applicability. Moreover, a trade-off was observed between reduced false negatives and increased false positives and fragmentations (Frag). Future work will focus on alleviating computational bottlenecks through candidate reduction, approximated distance correlation, and lightweight batch processing, as well as improving metric balance via conservative reactivation, low-confidence suppression, and fragmented track merging. Additionally, the evaluation will be extended to diverse sequences featuring varying crowd densities, occlusion patterns, and scene dynamics to systematically assess the generalizability of the proposed approach.

REFERENCES

- [1] C.-H. Seo, C. Q. Nguyen, D.-W. Heo, and S.-R. Choi, "Analysis of trends in multiple object tracking technology," *Journal of Institute of Control, Robotics and Systems (in Korean)*, vol. 30, no. 2, pp. 13–24, Feb. 2024.
- [2] H.-H. Nguyen, D. N.-N. Tran, L. H. Pham, T. H.-P. Tran, and J. W. Jeon, "City-scale multi-camera tracking of vehicles using YOLOv9 and ByteTrack," *Proceedings of the 24th International Conference on Control, Automation and Systems (ICCAS)*, Jeju, Korea, pp. 583–587, Oct.–Nov. 2024.
- [3] C. Y. Kyun and S. N. Yu, "Enhancing user convenience in virtual reality-based robotic control systems: utilizing catmull splines, hand tracking technology, and eye tracking technology," *Journal of Institute of Control, Robotics and Systems (in Korean)*, vol. 30, no. 9, pp. 1004–14, Sep. 2024.
- [4] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, USA, pp. 2544–2550, Jun. 2010.
- [5] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, USA, pp. 6931–6939, Jul. 2017.
- [6] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional Siamese networks for object tracking," *Proceedings of European Conference on Computer Vision (ECCV) Workshops*, Amsterdam, Netherlands, pp. 850–865, Oct. 2016.
- [7] Z. Chen, B. Zhong, G. Li, S. Zhang, and R. Ji, "Transformer tracking," *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, USA, pp. 8126–8135, Jun. 2021.
- [8] S. Hassan, G. Mujtaba, A. Rajput, and N. Fatima, "Multi-object tracking: A systematic literature review," *Multimedia Tools and Applications*, vol. 83, no. 14, pp. 43439–43492, 2024.
- [9] W. Luo, X. Xing, X. Zhang, and P. Li, "Multiple object tracking: A survey," arXiv preprint arXiv:2003.09003, 2020.
- [10] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Phoenix, USA, pp. 3464–3468, Sep. 2016.
- [11] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Beijing, China, pp. 3645–3649, Sep. 2017.
- [12] J. Zhang, X. Sun, S. Jin, et al., "BoT-SORT: Robust associations for multi-pedestrian tracking," *Proceedings of European Conference on Computer Vision (ECCV) Workshops*, Tel Aviv, Israel, pp. 1–18, Oct. 2022.

- [13] Y. Zhang, P. Sun, Y. Jiang, et al., "ByteTrack: Multi-object tracking by associating every detection box," *Proceedings of European Conference on Computer Vision (ECCV)*, Tel Aviv, Israel, pp. 1–21, Oct. 2022.
- [14] Y. Xu, Y. Ban, S. Peng, and A. M. Baskurt, "StrongSORT: Make DeepSORT great again," arXiv preprint arXiv:2202.13514, 2022.
- [15] J. Cao, J. Weng, R. Voirin, Z. Guo, and T. Zhao, "OC-SORT: Rethinking SORT for robust multi-object tracking," *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, pp. 9636–9646, Oct. 2023.
- [16] M. Yang, G. Han, B. Yan, W. Zhang, J. Qi, H. Lu, and D. Wang, "Hybrid-SORT: Weak cues matter for online multi-object tracking," *Proceedings of AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, pp. 6504–6512, 2024.
- [17] K. Yi, K. Luo, X. Luo, J. Huang, H. Wu, R. Hu, and W. Hao, "UCMCTrack: Multi-object tracking with uniform camera motion compensation," *Proceedings of AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, pp. 6702–6710, 2024.
- [18] Y. Jiang, W. Zhao, Y. Gao, X. Niu, J. Li, and X. Gao, "WM-SORT: Modeling multi-object tracking motion patterns in real world," *Neurocomputing*, vol. 638, p. 130182, 2025.
- [19] L. Chen and R. Tang, "ESC-SORT: Enhanced similarity costs lead to an improved multi-object tracker," *Proceedings of the 17th International Conference on Digital Image Processing (ICDIP 2025)*, vol. 13709, pp. 21–30, 2025.
- [20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Proceedings of NeurIPS*, 2015.
- [21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, pp. 779–788, Jun. 2016.
- [22] MOTChallenge, "MOT17 Challenge," <http://motchallenge.net/data/MOT17/>.
- [23] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing dependence by correlation of distances," *Annals of Statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.
- [24] G. J. Székely and M. L. Rizzo, "Brownian distance covariance," *Annals of Applied Statistics*, vol. 3, no. 4, pp. 1236–1265, Dec. 2009.
- [25] X. Huo and G. J. Székely, "Fast computing for distance covariance," *Technometrics*, vol. 58, no. 4, pp. 435–447, Oct. 2016.
- [26] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *Proceedings of European Conference on Computer Vision (ECCV)*, 2020.
- [27] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," *Proceedings of ICLR*, 2021.
- [28] L. Rakai, H. Song, S. Sun, W. Zhang, and Y. Yang, "Data association in multiple object tracking: A survey of recent techniques," *Expert Systems with Applications*, vol. 192, p. 116300, Apr. 2022.
- [29] R. E. Kalman, "A new approach to linear filtering and prediction problems," *ASME Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, Mar. 1960.
- [30] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation: Theory, Algorithms, and Software*, Wiley, New York, USA, 2001.
- [31] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "MARS: A video benchmark for large-scale person re-identification," *Proceedings of European Conference on Computer Vision (ECCV)*, Amsterdam, Netherlands, pp. 868–884, Oct. 2016.
- [32] S. Zagoruyko and N. Komodakis, "Wide residual networks," *Proceedings of BMVC*, York, UK, pp. 87.1–87.12, Sep. 2016.
- [33] N. Wojke and A. Bewley, "Deep cosine metric learning for person re-identification," *Proceedings of IEEE WACV*, Lake Tahoe, USA, pp. 748–756, Mar. 2018.
- [34] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1–2, pp. 83–97, Mar. 1955.
- [35] Papers with Code, "MOT17 Benchmark (Multi-Object Tracking)," <https://paperswithcode.com/dataset/mot17>.
- [36] K. Bernardin and R. Stiefelhausen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *EURASIP Journal on Image and Video Processing*, 2008(1):246309, 2008.



김 유 진

2021년 인하공업전문대학 컴퓨터시스템 및 컴퓨터시스템공학과 졸업(학사). 2023년~현재 한양대학교 대학원 인공지능융합학과 석사와 정 재학 중. 관심분야는 컴퓨터비전, 딥러닝, 알고리즘.



강 경 태

1999년 서울대학교 컴퓨터 과학 공학 졸업 (학사). 2001년, 2007년 서울대학교 전기 공학 및 컴퓨터 과학과 석사, 박사 졸업. 2008년~2010년 일리노이 어배너-شم페인 대학 박사 후 연구원. 2011년~현재 한양대 컴퓨터공학과 교수. 관심 분야는 운영 체제, 모바일 시스템, 분산 시스템 및 실시간 임베디드 시스템.



권 보 경

2010년 University of Pittsburgh, Biostatistics, Ph.D. 2023년~현재 한양대학교 ERICA 인공지능학과 부교수. (전) 삼성전자 선임, (전) Hewlett-Packard 책임. 관심 분야는 기계학습, 컴퓨터 비전, Statistical Computing, Optimization and Algorithms.