Journal of Imaging

MDPI

*Article*

# Reliability-Aware Semantic Gating with Retrieval-Augmented Association for Online Multi-Object Tracking Under Industrial Low-Altitude Proxy Conditions

**Rongzuo Guo * and Yuhang He**

School of Computer Science, Sichuan Normal University, Chengdu 610068, China; 20241393048@stu.sicnu.edu.cn
* Correspondence: gyz000001@163.com

**Abstract**

Industrial low-altitude multi-object tracking is challenged by overhead viewpoints, dense small targets, mutual occlusion, homogeneous appearances, and domain-shifted backgrounds. Vision-language pre-trained models provide useful semantic priors, but their category-level representations can mislead instance-level association when many same-class targets are densely distributed. This study develops an OC-SORT-based framework with reliability-aware semantic gating and retrieval-augmented trajectory memory. On the MOT17 reporting split (MOT17-09/10/11/13) with FRCNN detections, the FullModel achieves 46.41% MOTA, 40.00% IDF1, 34.542% HOTA, 26.686% AssA, and 382 identity switches for seed 42. Relative to fixed-weight CLIP+FAISS, it reduces identity switches and fragmentations while trading 0.84 MOTA points for a 0.68-point IDF1 gain; exact reruns with three seeds reproduce the same MOTA, IDF1, and IDSw totals. The VisDrone-derived subset is used only as an industrial low-altitude proxy stress test: its high precision coexists with very low recall, so the result exposes domain-shift limitations, rather than demonstrating deployment-level industrial performance.

**Keywords:** multi-object tracking (MOT); observation-centric SORT (OC-SORT); contrastive language-image pre-training (CLIP); semantic noise; dynamic semantic gating; retrieval-augmented data association; industrial low-altitude proxy conditions

## 1. Introduction

Multi-object tracking (MOT) aims to simultaneously estimate the locations of multiple targets and maintain their identity consistency in consecutive video frames [1,2]. It is a fundamental component of visual perception systems such as intelligent surveillance, industrial inspection, warehouse logistics, and autonomous driving. With the increasing demand for industrial automation and intelligent safety supervision, stable, real-time, and identity-preserving MOT methods in complex scenarios have become highly important. However, in industrially relevant environments such as factory workshops, logistics warehouses, and low-altitude inspection scenes, dense target distributions, long-term occlusions, highly similar appearances, and drastic viewpoint changes often occur simultaneously. These factors make reliable identity association challenging for existing tracking algorithms.

Compared with standard pedestrian surveillance benchmarks, industrial low-altitude scenes exhibit several distinctive challenges. First, overhead or oblique viewpoints make target scales highly unstable and reduce the discriminative details of human bodies, vehicles, and equipment. Second, workers often wear similar uniforms, helmets, or reflective

Check for updates

Academic Editor: Pierre Gouton

Received: 6 July 2026
Revised: 2 September 2026
Accepted: 8 September 2026
Published: 16 September 2026

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

vests, leading to strong appearance homogeneity. Third, industrial facilities such as forklifts, conveyors, pallets, shelves, and safety barriers frequently occlude moving targets and generate structured background interference. Fourth, low-altitude unmanned aerial vehicle (UAV) platforms introduce viewpoint changes and non-stationary observation geometry, making constant-velocity motion prediction less reliable. These factors make industrial MOT different from general pedestrian tracking, where identity association can often rely on relatively stable appearance and motion cues. Therefore, industrial low-altitude MOT requires not only general tracking robustness, but also scene-adaptive semantic priors and reliability-aware cue regulation.

In recent years, tracking-by-detection methods have achieved substantial progress in MOT. Simple Online and Real-time Tracking (SORT) [3], Deep SORT (DeepSORT) [4], Observation-Centric SORT (OC-SORT) [5], ByteTrack [6], StrongSORT [7], and Deep OC-SORT [8] have improved online tracking performance from different perspectives, including motion prediction, appearance representation, utilization of low-confidence detections, and trajectory recovery. Meanwhile, You Only Look Once (YOLO)-series detectors have provided efficient detection front-ends for real-time MOT [9–11]. Re-identification (ReID) features, pose alignment, and pairwise similarity optimization have also played important roles in distinguishing occluded targets and visually similar objects [12–15]. More recently, contrastive language-image pre-training (CLIP) [16] models have been introduced into open-vocabulary detection, cross-domain recognition, and MOT tasks, providing new semantic priors for visual target association [17,18].

However, direct transfer from standard surveillance benchmarks to industrial low-altitude scenes remains difficult. To quantitatively characterize this degradation, a controlled reproduction is conducted on the VisDrone2019-MOT-val industrial proxy subset [19] under the same YOLOv8s detections. Multi-Object Tracking Accuracy (MOTA), Identity F1 Score (IDF1), recall, and precision are used as evaluation metrics. OC-SORT obtains −22.3% MOTA, 7.2% IDF1, 16.4% recall, and 45.2% precision, while ByteTrack obtains −21.8% MOTA, 8.1% IDF1, 17.6% recall, and 46.5% precision. The negative MOTA values indicate that false positives, missed detections, and identity switches (IDSw) dominate the correct tracking outputs, whereas the low IDF1 values reveal severe identity inconsistency. These results show that mainstream trackers suffer from substantial performance degradation under overhead viewpoints, dense small targets, frequent occlusion, and domain-shifted backgrounds.

It should be noted that the VisDrone2019-MOT-val industrial proxy subset is not a complete factory-floor MOT benchmark. It is used in this paper only as a proxy stress test for overhead viewpoints, dense small targets, and complex backgrounds. Therefore, the results should not be interpreted as proof of deployment-level industrial performance; they mainly reveal the failure modes and transfer limitations that must be addressed before real factory-floor application.

Existing studies mainly focus on the open-category generalization ability of vision-language models, while insufficient attention has been paid to the potential negative effects of their category-level semantic features in instance-level association among dense objects of the same class. In industrial scenarios, workers wearing uniform clothing, devices of similar models, and dense spatial layouts are common. If CLIP semantic features are directly fused with fixed weights, category consistency may be enhanced while individual differences are weakened, thereby increasing the risk of IDSw.

To address this issue, this paper investigates the reliability of vision-language semantic features in instance-level MOT association and proposes a reliability-aware semantic gating and retrieval-augmented MOT method. Based on OC-SORT, the proposed method analyzes the applicability boundary of CLIP category-level semantic features from the perspective

of trajectory-detection candidate pairs. A semantic noise event is defined as the situation in which the semantic similarity between an incorrect same-class candidate detection and the trajectory prototype is no lower than that between the true matched detection and the trajectory prototype. On this basis, several offline metrics are constructed, including semantic discriminative margin, Semantic Noise Rate (SNR), Semantic Noise Magnitude (SNM), and Semantic Collapse Index (SCI), to characterize the misleading risk of CLIP semantic features in instance-level association.

Furthermore, a candidate-pair-level dynamic semantic gating (DSG) mechanism is designed. It estimates semantic reliability using detection confidence, spatial visibility, motion consistency, and target density, and adaptively regulates the CLIP visual semantic term and dictionary-retrieval term. In addition, retrieval-augmented semantic memory is constructed to enhance semantic stability under occlusion recovery, appearance degradation, and homogeneous target association. In the current implementation, the semantic dictionary is used as an auxiliary retrieval reference, rather than as a complete industrial knowledge base.

The main contributions of this paper are summarized as follows.

(1) Semantic noise quantification for CLIP-assisted MOT. This paper analyzes the negative effect of CLIP category-level semantics in instance-level trajectory association. From the perspective of trajectory-detection candidate pairs, semantic noise events are defined, and semantic discriminative margin, Semantic Noise Rate (SNR), Semantic Noise Magnitude (SNM), and Semantic Collapse Index (SCI) are introduced to quantify the risk of same-class but different-identity targets becoming over-clustered in the CLIP feature space.

(2) Reliability-aware dynamic semantic gating. A candidate-pair-level reliability estimation mechanism is proposed to regulate semantic costs during data association. Detection confidence, spatial visibility, motion consistency, and target density are used as observable reliability cues. The resulting gating coefficient dynamically suppresses CLIP semantic cost and dictionary-retrieval cost under unreliable or highly ambiguous association states.

(3) Retrieval-augmented trajectory semantic memory. A retrieval-augmented semantic memory is constructed to provide temporally smoothed semantic references for trajectory association. Instead of using CLIP features as fixed-weight matching cues, the proposed method treats semantic information as a reliability-constrained auxiliary signal, which improves robustness against semantic noise and supports preliminary transfer to industrial low-altitude proxy scenes.

The remainder of this paper is organized as follows. Section 2 reviews related studies on object detection, multi-object tracking, occlusion and similar-appearance handling, vision-language models in MOT, and industrial low-altitude tracking scenarios. Section 3 presents the proposed retrieval-augmented reliability-aware semantic association model. Section 4 evaluates the proposed method from the perspectives of comparison with mainstream trackers, semantic noise analysis, module ablation, reliability-grouped analysis, cross-scenario transfer, runtime efficiency, and failure-case analysis. Section 5 discusses the results. Section 6 concludes the paper and discusses future research directions.

## 2. Related Work

### *2.1. Object-Detection Algorithms*

Object detection is an essential front-end for MOT. Its accuracy, recall, and real-time performance directly affect subsequent trajectory association. In recent years, object detection has evolved from traditional closed-set detection to real-time detection, open-vocabulary

detection, and vision-language joint perception. Zou et al. [20] systematically reviewed the development of object detection from handcrafted features and convolutional neural networks to Transformer-based detectors, and pointed out that small objects, occlusion, long-tailed categories, and complex backgrounds remain major factors limiting detection performance. Meanwhile, YOLO-series methods have become widely used detection front-ends in online MOT, due to their favorable speed–accuracy trade-off [9–11]. Real-time Transformer detectors such as RT-DETR further indicate that global attention modeling provides new structural choices for real-time detection [21,22].

Although existing detectors have achieved high performance in standard scenarios, industrial scenarios still suffer from missed detections and localization errors caused by occlusion, low illumination, small-scale objects, and overhead viewpoints. Detection errors are further amplified during tracking, leading to trajectory fragmentation, false association, and identity switches. Therefore, improving detector performance alone is insufficient to fully solve tracking instability in complex scenarios. More robust multi-cue fusion and reliability regulation mechanisms are still needed in the data association stage.

### 2.2. Multi-Object Tracking Algorithms

Current mainstream MOT methods are mainly based on the tracking-by-detection paradigm [1,2]. Candidate bounding boxes are first generated by a detector in each frame, and then data association is performed to maintain identity continuity. SORT [3] establishes a simple and efficient online tracking framework using Kalman filtering and Hungarian matching. DeepSORT [4] introduces ReID appearance features to improve the discrimination of visually similar targets. ByteTrack [6] effectively exploits low-confidence detections to improve target recall in occlusion and missed-detection scenarios. OC-SORT [5] improves motion prediction and trajectory association from an observation-centric perspective, enhancing trajectory stability under occlusion. StrongSORT [7] and Deep OC-SORT [8] further integrate appearance modeling, trajectory recovery, and multi-cue association strategies to advance online MOT performance.

Other representative tracking-by-detection methods, such as BoT-SORT [23], FairMOT [24], Tracktor [25], and CenterTrack [26], also improve MOT from the perspectives of robust association, joint detection–re-identification learning, regression-based tracking, and point-based tracking. In addition, Transformer-based and end-to-end tracking methods such as TrackFormer [27], MOTR [28], and MOTRv2 [29] attempt to model temporal association through object queries and integrate detection and association into a unified framework. Dataset studies such as DanceTrack [30] and SportsMOT [31] further demonstrate that in scenarios with highly uniform appearances and complex motion patterns, relying solely on appearance features is insufficient to maintain stable identity association.

Overall, existing methods have achieved important progress in motion modeling and appearance representation, but how to dynamically regulate the reliability of different association cues according to target states remains an open problem in complex industrial scenarios.

### 2.3. Occlusion and Similar-Appearance Handling

Occlusion and similar appearance are key factors affecting identity consistency in MOT. Existing studies on occlusion mainly follow two directions: improving detection robustness and optimizing trajectory association. One class of methods enhances the detector's perception of occluded targets through data augmentation, occlusion simulation, or network-structure improvement. Another class uses motion prediction, trajectory completion, and re-identification mechanisms to maintain identity continuity after short-term

occlusion. These methods alleviate trajectory fragmentation to some extent, but substantial uncertainty remains under long-term full occlusion or dense target interaction.

For distinguishing targets with similar appearances, ReID features are commonly used as identity representation cues. Ye et al. [12] provided a systematic review of deep person re-identification. FastReID [13] offers a comprehensive training and evaluation framework for general instance re-identification. Pose-guided feature alignment [14] and Circle Loss [15] improve feature discriminability from the perspectives of local alignment and pairwise similarity optimization, respectively. However, targets in industrial scenarios often exhibit highly uniform appearances, such as workers wearing similar uniforms, equipment of similar models, and repeatedly appearing safety facilities. In such cases, traditional appearance features are prone to homogenization, and relying solely on ReID or local visual texture is insufficient to distinguish different instances. Therefore, higher-level semantic priors should be introduced, while the reliability of semantic information must be constrained according to target states.

### *2.4. Applications and Challenges of CLIP in MOT*

CLIP [16] learns strong cross-modal semantic alignment through large-scale image–text contrastive learning, promoting the development of open-vocabulary detection, open-category recognition, and general visual perception. OpenCLIP [32] and SigLIP [33] further verify the scalability and stability of image–text contrastive learning in large-scale pre-training. ViLD [34], GLIP [35], RegionCLIP [36], Detic [37], and Grounding DINO [38] improve open-vocabulary object perception through knowledge distillation, phrase grounding, region-level image–text pre-training, and text-guided detection. In addition, general visual models such as BLIP-2 [39] and Segment Anything [40] demonstrate that large-scale pre-trained models can provide richer semantic representations for complex vision tasks.

In MOT, vision-language models are mainly used to enhance open-category recognition or cross-domain generalization. OVTrack [17] shows that CLIP semantic features can provide effective complementary information for open-vocabulary MOT. However, CLIP features intrinsically emphasize category-level semantic consistency, whereas identity preservation in MOT requires distinguishing different instances within the same category. When targets are densely distributed and highly similar in appearance, category-level semantic features may pull different individuals too close in the feature space, thereby weakening instance-level discrimination. Existing studies have insufficiently discussed this negative effect, and lack mechanisms to dynamically adjust semantic weights according to detection confidence, occlusion degree, and motion consistency. Therefore, how to exploit vision-language semantic priors while suppressing their misleading effects under low-reliability association states is the core problem addressed in this paper.

### *2.5. Multi-Object Tracking in Industrial and Low-Altitude Scenarios*

Industrial-related MOT studies can be roughly divided into low-altitude UAV tracking, traffic-surveillance tracking, and factory or warehouse monitoring. VisDrone [19] and UAVDT [41] focus on UAV-view detection and tracking, where small targets, camera motion, and drastic viewpoint changes are dominant challenges. UA-DETRAC [42] and BDD100K [43] provide large-scale traffic scenarios with vehicle density, illumination changes, and long-term occlusion. Compared with these settings, factory and warehouse scenarios further introduce domain-specific difficulties, including uniform workwear, repeated equipment patterns, safety-facility occlusions, and structured production layouts. These factors make identity association more difficult than in general pedestrian or traffic benchmarks.

Most existing public MOT benchmarks do not fully cover factory-floor production scenes. UAV-view datasets such as VisDrone and UAVDT emphasize small targets, camera motion, and viewpoint changes, while traffic datasets such as UA-DETRAC and BDD100K mainly focus on road vehicles. Factory and warehouse scenarios may additionally contain uniform workwear, repeated equipment layouts, conveyor facilities, pallets, and safety barriers. Due to the limited availability of public factory-floor MOT benchmarks, this paper uses a VisDrone-derived industrial proxy subset only for preliminary cross-scenario transfer analysis. This subset is intended to approximate several visual characteristics of industrial low-altitude scenes, including overhead viewpoints, dense small targets, and complex backgrounds. It should not be regarded as a complete factory-floor MOT benchmark.

## 3. Materials and Methods

### 3.1. Overall Framework

To improve identity preservation in complex industrial scenarios, this paper constructs a retrieval-augmented reliability-aware semantic association model based on OC-SORT [5]. The proposed model integrates geometric motion, appearance representation, CLIP visual semantic features, and dictionary-retrieval semantic representation into the data association cost. A candidate-pair-level dynamic gating mechanism is used to regulate the contribution of semantic terms under different association states. The overall framework is shown in Figure 1.

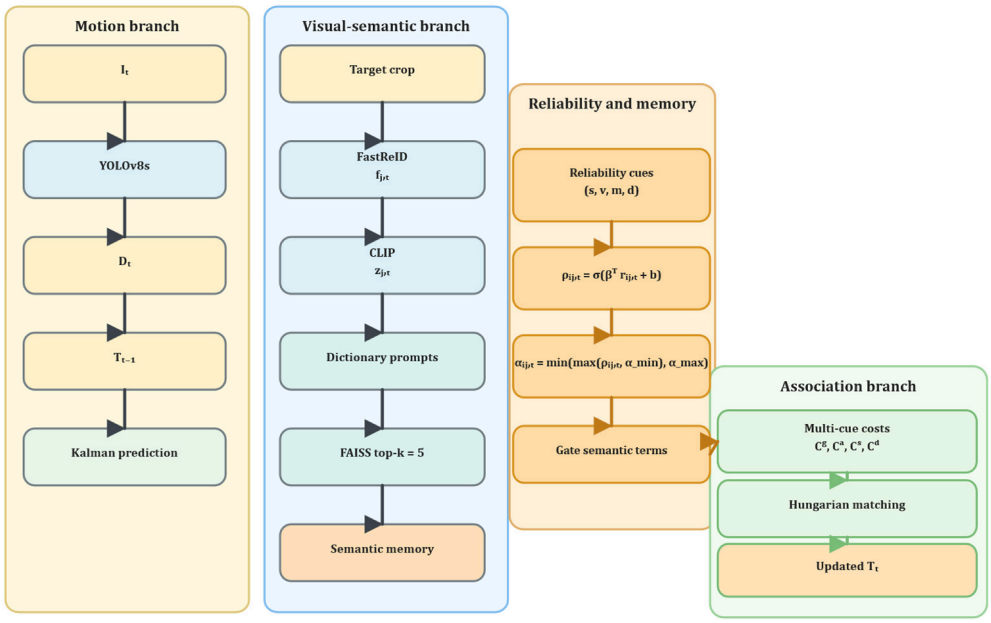


**Figure 1.** Vector schematic of the proposed reliability-aware semantic gating and retrieval-augmented MOT framework.

In the proposed framework, the detector first generates candidate observations in the current frame. OC-SORT predicts historical trajectory states using Kalman filtering [44]. For each retained detection crop, FastReID extracts a 2048-dimensional appearance descriptor and CLIP-ViT-B/32 extracts a 512-dimensional L2-normalized visual-semantic feature. The CLIP feature is queried against the text-prompt dictionary through FAISS IndexFlatIP; the top-k = 5 inner-product neighbors are used to form a sparse dictionary response before softmax normalization. For each trajectory-detection candidate pair, the model estimates semantic reliability from detection confidence, spatial visibility, motion consistency, and target density, maps the score to a dynamic gating coefficient, and fuses geometric cost,

FastReID appearance cost, CLIP semantic cost, and dictionary retrieval cost under reliability constraints. Matching is finally solved using the Hungarian algorithm [45].

### 3.2. Quantitative Definitions of Semantic Noise and Semantic Collapse

To avoid treating semantic noise and semantic collapse as merely empirical descriptions, this paper provides computable definitions from the perspective of trajectory-detection candidate association. Let $F_v$ denote the frozen CLIP visual encoder and let $(I_a, B_a)$ extract the target crop at bounding box $B_a$ from image $I_a$. After L2 normalization, the visual-semantic features $z_a$ and $z_b$ in $R^{512}$ of observations a and b are defined as follows:

$$\begin{aligned} z_a &= F_v(crop(I_a, B_a))/||F_v(crop(I_a, B_a))||_2, \\ z_b &= F_v(crop(I_b, B_b))/||F_v(crop(I_b, B_b))||_2, \end{aligned} \tag{1}$$

The CLIP semantic similarity between them is defined as

$$\mathrm{s(a,b)} = \mathrm{z_a^T z_b}, \tag{2}$$

Let y(·) denote the identity label and c(·) denote the category label. The positive pair set P and the same-class negative pair set Nc are defined as

$$\mathrm{P} = \{(\mathrm{a,b}) \mid \mathrm{y_a} = \mathrm{y_b}, \mathrm{a} \neq \mathrm{b}\}, \tag{3}$$

$$\mathrm{N_c} = \{(\mathrm{a,b}) \mid \mathrm{y_a} \neq \mathrm{y_b}, \mathrm{c_a} = \mathrm{c_b}, (\mathrm{a,b}) \in \mathrm{G}\}, \tag{4}$$

where $G$ is the candidate-association gating set obtained from spatial overlap or motion-prediction constraints. The operator denotes intersection over union when spatial overlap is used. Pairs in $P$ correspond to observations of the same identity at different moments, whereas pairs in $N_c$ correspond to candidate observations with the same category but different identities. Based on these sets, the semantic discriminative margin is defined as

$$\Delta_{\mathrm{sem}} = \mathbb{E}_{(\mathrm{a,b})\in\mathcal{P}}[\mathrm{s(a,b)}] - \mathbb{E}_{(\mathrm{a,b})\in\mathcal{N}_c}[\mathrm{s(a,b)}], \tag{5}$$

A larger Δsem indicates that CLIP semantic features can better distinguish the same identity from different identities within the same class. When Δsem approaches zero or becomes negative, different same-class identities are excessively close in the CLIP feature space, indicating weak instance-level discriminability.

For trajectory i, consider a frame t for which a ground-truth match is available. Let $j_{i,t}^{+}$ denote the true matched detection and let $N_{i,t}^{-}$ contain the incorrect same-class candidate detections that pass the candidate gate. The L2-normalized historical semantic prototype before frame t is defined as

$$\bar{\mathrm{z}}_{\mathrm{i,t-1}} = \frac{\sum_{\tau\in\mathrm{H_i^{t-1}}} \mathrm{z_{i,\tau}}}{\left\|\sum_{\tau\in\mathrm{H_i^{t-1}}} \mathrm{z_{i,\tau}}\right\|_2 + \epsilon}, \tag{6}$$

In Equation (6), $\mathbf{H}_{i,t-1}$ denotes the set of valid observation frames associated with trajectory i before frame t. For each historical frame τ in $\mathbf{H}_{i,t-1}$, $\mathbf{z}_{i,\tau}$ in $R^{512}$ denotes the L2-normalized CLIP visual-semantic feature extracted from trajectory i's observation at frame τ. The scalar $\varepsilon > 0$ is a small numerical constant used to prevent division by zero when the averaged feature vector has negligible norm. The numerator averages the historical feature vectors and the denominator normalizes this average; therefore, $\mathbf{z}_{i,\bar{t}-1}$ is a unit-norm semantic prototype. The true-match similarity is

$$\mathrm{s_{i,t}^{+}} = \bar{\mathrm{z}}_{\mathrm{i,t-1}}^{\mathrm{T}} \mathrm{z_{j_{i,t}^{+},t}}, \tag{7}$$

Equation (8) uses the hardest negative candidate to quantify the most misleading same-class alternative. Here $\mathbf{N}_{i,t}^{-}$ denotes the set of incorrect same-class detections that satisfy the candidate gate, and $\mathbf{z}_{j,t}$ denotes the L2-normalized CLIP feature of candidate detection j in frame t. The maximum selects the candidate with the greatest similarity to the historical prototype, so $s_{i,t}^{-}$ represents the strongest competing semantic match; $s_{i,t}^{+}$ is the similarity to the true matched detection defined in Equation (7).

$$s_{i,t}^{-} = \max_{j \in N_{i,t}^{-}} \bar{z}_{i,t-1}^{T} z_{j,t}, \tag{8}$$

If si,t− ≥ si,t+, the incorrect same-class candidate detection is closer to the historical semantic prototype of the trajectory than the true matched detection in the CLIP semantic space. In this case, if the CLIP semantic cost is fused with a fixed weight, the semantic term may mislead data association. This situation is defined as a semantic noise event.

Let $A$ denote the set of all association samples (i,t) for which a ground-truth matched detection exists and at least one incorrect candidate of the same category passes the candidate gate. In Equation (9), $\mathbb{I}(\cdot)$ is the indicator function: it equals 1 when the condition inside the parentheses is true and 0 otherwise. Thus, SNR is the fraction of samples in $A$ for which the hardest incorrect candidate is at least as similar as the true match, as measured by $s_{i,t}^{-} \geq s_{i,t}^{+}$. The Semantic Noise Rate (SNR) is defined as

$$R_{SN} = \frac{1}{|A|} \sum_{(i,t) \in A} 1\left(s_{i,t}^{-} \geq s_{i,t}^{+}\right), \tag{9}$$

A larger SNR means that CLIP semantic features more frequently rank an incorrect same-class candidate above the true match during candidate association.

To further measure the strength of semantic noise, the Semantic Noise Magnitude (SNM) is defined as

$$M_{SN} = \frac{1}{|A|} \sum_{(i,t) \in A} \max\left(0, s_{i,t}^{-} - s_{i,t}^{+}\right), \tag{10}$$

To quantify semantic collapse, this paper defines the proportion of high-similarity pairs among same-class different-identity targets, namely the Semantic Collapse Index (SCI):

$$C_{SC}(\tau) = \frac{1}{|N_c|} \sum_{(a,b) \in N_c} 1(s(a,b) > \tau), \tag{11}$$

In the experiments, $\tau = 0.80$ is used by default. $\Delta_{sem}$, $SNR$, $SNM$, and $SCI_{\tau}$ all depend on ground-truth identity labels. They are used only for offline analysis of CLIP semantic reliability in instance-level association, and do not participate in online tracking.

### 3.3. Candidate-Pair-Level Reliability-Aware Semantic Gating

CLIP features tend to represent category-level semantics. In dense same-class target scenarios, incorrect candidate detections may obtain semantic similarities close to, or even higher than, the true target. Since true identity labels cannot be directly observed during online tracking, semantic noise events cannot be explicitly identified. Therefore, this paper introduces a candidate-pair-level reliability estimation mechanism. Detection quality, spatial visibility, motion consistency, and target density are used as observable proxy variables to estimate the reliability of semantic cues in the association process.

For the candidate association pair formed by trajectory i and detection box j in frame t, the reliability vector is defined as

$$r_{ij,t} = \left[s_{j,t} \quad , v_{j,t} \quad , m_{ij,t}\right]^{\top}, \tag{12}$$

where sj,t ∈ [0, 1] denotes the normalized detection confidence of detection j, vj,t ∈ [0, 1] denotes its spatial visibility, and mij,t ∈ [0, 1] denotes the motion consistency between trajectory i and detection j.

The spatial visibility is computed from the overlap between the current detection and other detections in the same frame:

$$o_{j,t} = \min\left(1, \sum_{k \neq j} \frac{\left|B_{j,t} \cap B_{k,t}\right|}{\left|B_{j,t}\right| + \varepsilon}\right), \; v_{j,t} = 1 - o_{j,t}, \tag{13}$$

$$d_M^2\left(hatx_{i,t}, x_{j,t}\right) = \left(x_{j,t} - hatx_{i,t}\right)^T S_{i,t}^{-1}\left(x_{j,t} - hatx_{i,t}\right), \tag{14}$$

The motion consistency is defined based on the Mahalanobis distance between the predicted state of trajectory i and the observation state of detection j:

$$m_{ij,t} = \exp\left(-\frac{d_M^2\left(\hat{X}_{i,t}, X_{j,t}\right)}{\lambda_m}\right), \tag{15}$$

Equation (13) defines spatial visibility, Equation (14) gives the expanded Mahalanobis distance, and Equation (15) maps that distance to motion consistency. Here, $\widehat{x_{i,t}}$ and $x_{j,t}$ are four-dimensional observation vectors $\left[c_x, c_y, w, h\right]^T$ representing the predicted trajectory state and the current detection, respectively. $S_{i,t}$ in $R^{4\times4}$ is the innovation covariance matrix, and ${S_{i,t}}^{-1}$ is its inverse. The parameter $\lambda_m$ = 9.49 is the 95% chi-square scale for the four-dimensional motion gate. Equations (16)–(19) then define the reliability score, density terms, and dynamic semantic gate.

The basic reliability score of the candidate pair is defined as

$$\rho_{ij,t} = \sigma\left(\beta^T r_{ij,t} + b\right), \tag{16}$$

where $(u) = \frac{1}{1+(-u)}$ is the sigmoid function, $\beta = [0.40, 0.35, 0.25]^T$ weights detection confidence, spatial visibility, and motion consistency, respectively, and $b = -0.35$ is the reliability-score bias.

To further consider scene crowdedness, the target density of frame t is defined as

$$d_t = \min\left(\frac{N_t}{N_{max}}, 1\right), \tag{17}$$

where $N_t$ is the number of detections in frame t, and $N_{max}$ is the normalization upper bound for the number of targets. Finally, the dynamic semantic gating coefficient is computed as

$$\bar{s}_t = \frac{1}{N_t}\sum_{j=1}^{N_t} s_{j,t}, \tag{18}$$

where $\gamma_c = 0.70, \gamma_d = 0.70, \tau_{den} = 0.30, \alpha_{min} = 0.05, \alpha_{max} = 1.00$, and $N_{max} = 100$. The clipping operator is explicitly defined as $(x, a, b) = \{\{x, a\}, b\}$. The values were fixed on the parameter-selection split as manual starting values: $\beta$ assigns the largest weight to detection confidence, while $\gamma_c$, $\gamma_d$, and $\tau_{den}$ regulate semantic terms when confidence or density becomes unreliable. $N_{max}$ is a normalization cap for per-frame target density, not a learned estimate of dataset size. The coupled gamma grid in Section 4.5 assesses only $\gamma_c$ and $\gamma_d$; a complete independent sensitivity study of all parameters remains work for the future.

$$\alpha_{ij,t} = \mathrm{clip}\left(\rho_{ij,t}[1 + \gamma_c(\bar{s}_t - 0.5)][1 - \gamma_d \max(0, d_t - \tau_{den})\right], \alpha_{min}, \alpha_{max}), \tag{19}$$

### 3.4. Retrieval-Augmented Trajectory Semantic Memory

This paper constructs a semantic dictionary and trajectory-level semantic memory. The dictionary contains a general category layer and two domain-related prompt layers, and the exact prompt templates are documented in Appendix B. In the current experiments, the prompt ensemble is used as an auxiliary retrieval reference, rather than as a complete industrial knowledge base.

Let the image of frame t be It, and let the detection box be Bj,t. The target region is cropped and fed into the CLIP visual encoder [16] to obtain the normalized semantic feature:

$$z_{j,t} = \frac{F_v(\mathrm{crop}(I_t, B_{j,t}))}{\|F_v(\mathrm{crop}(I_t, B_{j,t}))\|_2}, \tag{20}$$

For an established trajectory i, its historical CLIP visual semantic prototype $z_bar_{i,t}$ is maintained. When trajectory i is successfully matched with detection j, the trajectory semantic prototype is updated using the exponential moving average with $\eta_z = 0.10$:

$$\bar{z}_{i,t} = \frac{(1-\eta_z)\bar{z}_{i,t-1} + \eta_z z_{j,t}}{\|(1-\eta_z)\bar{z}_{i,t-1} + \eta_z z_{j,t}\|_2}, \tag{21}$$

Let pk be the prompt corresponding to the k-th dictionary entry. Its text semantic prototype is defined as

$$e_k = \frac{F_t(p_k)}{\|F_t(p_k)\|_2}, \tag{22}$$

where Ft(·) denotes the CLIP text encoder. The dictionary semantic prototype matrix is denoted as

$$E = [e_1, e_2, \ldots, e_K]^T, \tag{23}$$

For detection target j, the response distribution between its visual semantic feature and the dictionary text prototypes is defined as

$$q_{j,t} = \mathrm{Softmax}\left(\frac{Ez_{j,t}}{\lambda_q}\right), \tag{24}$$

where $\lambda_q = 0.07$ is the temperature coefficient. In practice, the dense dictionary matrix is stored in a FAISS IndexFlatIP index; only the top-k = 5 inner-product neighbors are retrieved for each detection, and the softmax is applied to this retrieved candidate set, rather than to the entire dictionary.

$$\bar{q}_{i,t} = (1-\eta_q)\bar{q}_{i,t-1} + \eta_q q_{j,t}, \tag{25}$$

When trajectory i is successfully matched with detection j, the dictionary memory is updated with $\eta_q = 0.10$. This trajectory-level memory smooths single-frame semantic-response fluctuations and provides a stable auxiliary reference during occlusion recovery and appearance degradation.

### 3.5. Reliability-Constrained Multi-Cue Data Association

The core of online MOT is to establish reliable one-to-one correspondences between current-frame detections and historical trajectory states. Let the active trajectory set in frame t be

$$T_{t-1} = \left\{\tau_i\right\}_{i=1}^{M}, \tag{26}$$

and the detection set be

$$\mathcal{D}_t = \left\{d_j = (B_{j,t}, s_{j,t}, c_{j,t})\right\}_{j=1}^{N_t}, \tag{27}$$

For trajectory τi, Kalman filtering [44] is first used to predict its state and bounding box in the current frame, denoted as x̂i,t and B̂i,t, respectively. To reduce computational burden and false association risk caused by invalid candidate matches, a candidate gating mechanism is introduced before cost calculation. Only trajectory-detection pairs satisfying spatial overlap or motion constraints are retained:

$$G_t = \{(i,j) \mid \mathrm{IoU}(\hat{B}_{i,t}, B_{j,t}) \geq \theta_{iou} \text{ or } d_M(\hat{x}_{i,t}, x_{j,t}) \leq \theta_m\}, \tag{28}$$

For candidate pair (i,j) ∈ Gt, four types of cues are integrated to construct the association cost: geometric motion, appearance representation, CLIP visual semantics, and dictionary retrieval semantics. The geometric cost is defined as

$$C^g_{ij,t} = 1 - \mathrm{IoU}(\hat{B}_{i,t}, B_{j,t}), \tag{29}$$

The appearance cost is defined as

$$C^a_{ij,t} = 1 - \frac{\overline{f}^T_{i,t-1} f_{j,t}}{\|\overline{f}_{i,t-1}\|_2 \|f_{j,t}\|_2 + \varepsilon}, \tag{30}$$

The CLIP visual semantic cost is defined as

$$C^s_{ij,t} = 1 - \overline{z}^T_{i,t-1} z_{j,t}, \tag{31}$$

The dictionary retrieval cost is defined as

$$C^d_{ij,t} = 1 - \frac{\overline{q}^T_{i,t-1} q_{j,t}}{\|\overline{q}_{i,t-1}\|_2 \|q_{j,t}\|_2 + \varepsilon}, \tag{32}$$

Considering that CLIP visual semantics and dictionary retrieval semantics may suffer from semantic collapse in dense same-class target scenarios, this paper does not directly fuse semantic terms with fixed weights. Instead, the dynamic gating coefficient is introduced to regulate semantic-related costs in a reliability-aware manner. The final total cost of the candidate pair is defined as

$$C_{ij,t} = w_g C^g_{ij,t} + w_a C^a_{ij,t} + \alpha_{ij,t}\left(w_s C^s_{ij,t} + w_d C^d_{ij,t}\right), \tag{33}$$

In this paper, $w_g = 0.45, w_a = 0.30, w_s = 0.15$, and $w_d = 0.08$. The FastReID appearance feature $f_{j,t}$ is a 2048-dimensional descriptor and is updated by the exponential moving average in the same manner as the semantic prototype. $C_t = [\, C_{ij,t} \,]$ is the candidate-pair cost matrix. The Hungarian algorithm [45] solves the one-to-one assignment in Equation (34).

$$\Pi^*_t = \underset{\Pi_t}{argmin} \sum_{(i,j)\in\Pi_t} C_{ij,t}, \tag{34}$$

### 3.6. Online Algorithm Procedure

Algorithm 1 summarizes the reliability-aware semantic gating and retrieval-augmented association procedure. It accepts the current frame image, the current detection set, the historical trajectory set, and the semantic dictionary, and returns the updated trajectory set. Its core steps are detection acquisition, Kalman prediction, FastReID appearance extraction, CLIP feature extraction, FAISS top-k dictionary retrieval, candidate-pair reliability estimation, reliability-aware multi-cue cost construction, Hungarian matching, and trajectory-memory updating. For the reported MOT17 experiments, the supplied FRCNN detection set is used in place of detector inference.

**Algorithm 1:** Reliability-Aware Semantic Gating and Retrieval-Augmented MOT

**Input:** Current frame image $(I_t)$, detection set $(D_t)$, historical trajectory set $(T_{t-1})$, dual-level semantic dictionary $(E)$.

**Output:** Updated trajectory set $(\mathrm{T_t})$.

Detect objects $(\mathrm{I_t})$ and obtain detection boxes, categories, and confidence scores.
for each trajectory $(\tau_i \in \mathrm{T_{t-1}})$ do
    Predict the state $(\hat{x}_{i,t})$ and bounding box $(\hat{B}_{i,t})$ using Kalman filtering.
end for
for each detection box $(B_{j,t} \in \mathcal{D}_t)$ do
    Crop the target image patch; extract FastReID appearance feature $f_{j,t}$ and CLIP visual feature $z_{j,t}$.
    Search the FAISS IndexFlatIP dictionary, retrieve top-k = 5 prompts, and compute $q_{j,t}$.
    Compute the spatial visibility $(v_{j,t})$.
end for
Compute the average detection confidence $(\bar{s}_t)$ and target density $(d_t)$.
Initialize the total cost matrix $(C_t)$.
for each trajectory–detection pair $((i,j))$ do
    if $((i,j) \notin \mathcal{G}_t)$ then
        $(C_{ij,t} \leftarrow +\infty)$.
        continue
    end if
    Compute the geometric cost $\left(C^{g}_{ij,t}\right)$.
    Compute the FastReID appearance cost $C^{a}_{ij,t}$.
    Compute the CLIP semantic cost $\left(C^{s}_{ij,t}\right)$.
    Compute the dictionary retrieval cost $C^{d}_{ij,t}$ from the retrieved FAISS response.
    Construct the reliability vector $(r_{ij,t} = [s_{j,t}, v_{j,t}, m_{ij,t}]^{\top})$.
    Compute the basic reliability score $(\rho_{ij,t})$.
    Compute the semantic gating coefficient $(\alpha_{ij,t})$.
    Compute the final association cost:
    $\left(C_{ij,t} = w_g C^{g}_{ij,t} + w_a C^{a}_{ij,t} + \alpha_{ij,t}\left(w_s C^{s}_{ij,t} + w_d C^{d}_{ij,t}\right)\right)$.
end for
Apply the Hungarian algorithm to $(C_t)$ and obtain the optimal matching result $\left(\Pi^{*}_{t}\right)$.
for each successfully matched pair $\left((i,j) \in \Pi^{*}_{t}\right)$ do
    Update the Kalman state, FastReID appearance prototype, CLIP semantic prototype, and dictionary semantic memory.
end for
Process unmatched trajectories and unmatched detections, remove expired trajectories, and initialize new trajectories.
return $(\mathrm{T_t})$.

## 4. Results

### 4.1. Experimental Settings

To verify the effectiveness of the proposed method in standard surveillance scenarios and to preliminarily examine its transfer limitations in low-altitude industrial-like scenes, experiments are conducted on the MOT17 training set [1] and the VisDrone2019-MOT-val industrial proxy subset [19]. For MOT17, a sequence-level split is used: MOT17-02, MOT17-04, and MOT17-05 are used only for parameter selection, while MOT17-09, MOT17-10, MOT17-11, and MOT17-13 are used for the reported validation results. No frame from the reporting split is used to tune the semantic-gating thresholds.

The VisDrone2019-MOT-val industrial proxy subset is used to approximate several visual characteristics of industrial low-altitude scenes, including overhead viewpoints, dense small targets, frequent occlusion, and complex backgrounds. Sequences are selected when they contain low-altitude UAV viewpoints, continuous multi-target interaction, and a sufficient number of annotated pedestrian/vehicle trajectories; empty or near-static clips are excluded. This subset is not a complete factory-floor MOT benchmark, and is not used to claim final deployment performance in real industrial scenarios. Instead, it is used as a stress test that exposes recall loss and detector-domain mismatch.

The evaluation follows the standard MOTChallenge protocol [1]. MOTA, IDF1, IDSw, MT, ML, precision, recall, and FPS are reported. HOTA [46] and AssA are additionally computed with the official TrackEval implementation on the same four-sequence MOT17 reporting split for FullModel and CLIP+FAISS.For instance-level reliability analysis of CLIP semantic features, the paper reports offline semantic discriminative margin, SNR, SNM, and SCI. These metrics require ground-truth identities and are used only for analysis, not online tracking.

All reported MOT17 results use the MOT17-09/10/11/13 reporting split, the supplied FRCNN detections, and seed 42, unless otherwise stated. The exact FullModel configuration is $w_g = 0.45, w_a = 0.30, w_s = 0.15$, and $w_d = 0.08$, $\gamma_c = 0.70, \gamma_d = 0.70$, and $\tau_{den} = 0.30$, as listed in Table 1. Experiments use an NVIDIA GeForce RTX 3090 GPU (NVIDIA Corporation, Santa Clara, CA, USA) with PyTorch 2.0. For the VisDrone proxy reproduction, the detector front end is Ultralytics YOLOv8s 8.0.20 with COCO-pretrained weights, a 640 × 640 input, confidence threshold 0.25, and NMS threshold 0.70. Tracker parameters are max_age = 30, min_hits = 3, and match_threshold = 0.50. FastReID uses a ResNet-50 backbone with 2048-dimensional descriptors. CLIP-ViT-B/32 uses a 224 × 224 input and 512-dimensional L2-normalized embeddings. FAISS [47] uses IndexFlatIP with normalized dictionary entries, top-k = 5 retrieval, and $\lambda_q = 0.07$.

Parameter-selection note: the listed values were fixed before the reported evaluation, and are not presented as globally optimal. Section 4.5 reports a coupled $\gamma_c/\gamma_d$ grid, but a one-at-a-time or joint sensitivity analysis for all critical parameters has not been completed.

**Table 1.** Hyperparameter settings of the proposed method.

| Parameter | Value | Description |
|---|---|---|
| $w_g$ | 0.45 | Base weight of geometric cost |
| $w_a$ | 0.30 | Appearance cost weight |
| $w_s$ | 0.15 | Base weight of CLIP visual semantic cost |
| $w_d$ | 0.08 | Base weight of dictionary retrieval cost |
| $\beta$ | [0.40, 0.35, 0.25] | Weights of detection confidence, spatial visibility, and motion consistency |
| $\tau_{den}$ | 0.30 | Density threshold in Equation (19) selected on the parameter split |
| $\gamma_d$ | 0.70 | Density suppression coefficient in Equation (19) |
| k | 5 | Number of retrieved nearest dictionary entries |
| match_threshold | 0.50 | Matching threshold of data association |
| max_age | 30 | Maximum unmatched frames for trajectory retention |
| min_hits | 3 | Minimum detection hits for trajectory confirmation |
| $N_{max}$ | 100 | Upper bound of target number normalization |
| CLIP input size | 224 × 224 | Input resolution of CLIP visual encoder |
| Detector input size | 640 × 640 | Input resolution of YOLOv8s detector |

**Table 1.** *Cont.*

| Parameter | Value | Description |
|---|---|---|
| $\gamma_c$ | 0.70 | Detection-confidence modulation coefficient in Equation (19) |
| $\alpha_{min}$ | 0.05 | Lower bound of the semantic gating coefficient |
| $\alpha_{max}$ | 1.00 | Upper bound of the semantic gating coefficient |
| b | −0.35 | Bias term of the reliability sigmoid in Equation (16) |
| $\lambda_m$ | 9.49 | Mahalanobis scale for motion consistency in Equation (15) |
| $\eta_z$ | 0.10 | EMA update rate for CLIP trajectory prototype |
| $\lambda_q$ | 0.07 | Temperature for dictionary-response softmax |
| $\eta_q$ | 0.10 | EMA update rate for dictionary semantic memory |
| YOLOv8s conf/NMS IoU | 0.25/0.70 | Detection confidence and non-maximum suppression thresholds |
| FastReID feature dimension | 2048 | ResNet-50 appearance descriptor dimension |
| CLIP feature dimension | 512 | ViT-B/32 visual-semantic descriptor dimension |

### *4.2. Comparison with Mainstream Methods*

To ensure fair comparison, the main comparison experiments include only tracking methods that can be reproduced under the same detection results, the same MOT17 reporting split, and the same evaluation protocol. The reporting split reduces leakage between parameter-selection and evaluation sequences, but all reported MOT17 results still come from the training set, rather than the official test server. Consequently, residual overfitting to the MOT17 training distribution cannot be excluded. Vision-language or end-to-end tracking methods such as OVTrack, GLEE, and MOTRv2 use different default detectors, training strategies, input resolutions, and model scales from this paper. Directly citing their original results would lead to unfair comparison. Therefore, results under different detectors and training settings are not mixed into the main performance table.

Table 2 reports the main comparison results on the MOT17 reporting split under FRCNN detections. The proposed method achieves 46.41% MOTA, 40.00% IDF1, 34.542% HOTA, 26.686% AssA, 382 ID switches, and 442 fragmentations. Compared with OC-SORT, the full model reduces both identity switches and fragmentations, but it does not dominate CLIP+FAISS retrieval in aggregate accuracy. This result is therefore better interpreted as an identity-stability trade-off than as a uniform performance gain.

**Table 2.** Main comparison on the MOT17 reporting split with FRCNN detections.

| Method | MOTA (%) | IDF1 (%) | HOTA (%) | AssA (%) | IDSw (Total) | Frag (Total) |
|---|---|---|---|---|---|---|
| ByteTrack | 39.64 | 49.85 | -- | -- | 810 | 297 |
| OC-SORT | 35.50 | 48.01 | -- | -- | 930 | 316 |
| CLIP + FAISS | 47.25 | 39.32 | 34.178 | 26.172 | 558 | 670 |
| FullModel (ours) | 46.41 | 40.00 | 34.542 | 26.686 | 382 | 442 |

Results are averaged over MOT17-09, MOT17-10, MOT17-11, and MOT17-13. IDSw and Frag are summed over the four sequences. FullModel uses CLIP and FAISS retrieval with the proposed DSG reliability gating. HOTA and AssA are computed with the official TrackEval implementation on the same four sequences for FullModel and CLIP+FAISS; -- indicates that the corresponding baseline was not evaluated with TrackEval.

All methods in Table 2 are evaluated on the same MOT17 reporting split with the supplied FRCNN detections. The HOTA and AssA entries use official TrackEval on seed 42 for FullModel and CLIP+FAISS; the MOTA, IDF1, and IDSw values for these two configurations were also reproduced identically with seeds 42, 100, and 2024. These reporting-split results should not be interpreted as an official MOT17 test-server evaluation.

### 4.3. Analysis of CLIP Semantic Noise and Semantic Collapse

Since the proposed semantic noise metrics require ground-truth identity labels and candidate-pair construction, the analysis focuses on same-ID similarity, same-class different-ID similarity, semantic discriminative margin, SNR, SNM, and SCI. The results are shown in Table 3.

**Table 3.** Quantitative analysis of CLIP semantic noise, including Semantic Noise Magnitude.

| Scenario | Same-ID Similarity | Same-Class Diff-ID Similarity | Discriminative Margin | SNR | SNM | SCI ($\tau$ = 0.8) |
|---|---|---|---|---|---|---|
| Low-density subset | 0.855 | 0.796 | 0.0590 | 12.7% | 0.018 | 18.3% |
| High-density MOT17-04 | 0.862 | 0.817 | 0.0451 | 21.4% | 0.031 | 32.6% |

In the low-density subset, the average same-ID similarity is 0.855, while the same-class different-ID similarity is 0.796, resulting in a discriminative margin of 0.0590. In the high-density MOT17-04 sequence, the same-class different-ID similarity increases to 0.817, and the discriminative margin decreases to 0.0451. Meanwhile, SNR increases from 12.7% to 21.4%, SNM increases from 0.018 to 0.031, and SCI increases from 18.3% to 32.6%. These results show that as target density increases, different identities belonging to the same category become more clustered in the CLIP semantic space. Consequently, CLIP category-level features may weaken instance-level discrimination when they are directly used for data association.

### 4.4. Unified Module Ablation Study

To verify the contribution of DSG, the ablation uses the same CLIP feature and FAISS retrieval pipeline under the MOT17 reporting split with FRCNN detections. Table 4 compares fixed semantic weighting without DSG against the FullModel with DSG.

**Table 4.** Ablation study on the MOT17 reporting split with FRCNN detections.

| Configuration | MOTA (%) | IDF1 (%) | HOTA (%) | AssA (%) | IDSw (Total) | Frag (Total) |
|---|---|---|---|---|---|---|
| CLIP + FAISS (fixed semantic weight, no DSG) | 47.25 | 39.32 | 34.178 | 26.172 | 558 | 670 |
| FullModel: CLIP + FAISS + DSG | 46.41 | 40.00 | 34.542 | 26.686 | 382 | 442 |

The exact configuration was rerun with seeds 42, 100, and 2024. All three runs produced identical four-sequence totals. FullModel reports IDSw 382.0 +/− 0.0, MOTA 46.41 +/− 0.00, and IDF1 40.00 +/− 0.00; CLIP+FAISS reports IDSw 558.0 +/− 0.0, MOTA 47.25 +/− 0.00, and IDF1 39.32 +/− 0.00. Because the paired differences are identical in all three runs, this zero-variance result is reported as deterministic multi-seed replication, not as evidence of stochastic significance; no conventional $p$-value claim is made.

Adding DSG reduces identity switches from 558 to 382 (a relative reduction of 31.5%) and fragmentations from 670 to 442 (34.0%), while changing MOTA by −0.84 percentage points and increasing IDF1 by 0.68 points. HOTA increases from 34.178 to 34.542 and AssA increases from 26.172 to 26.686 under the official TrackEval evaluation. This result shows that DSG primarily improves identity stability, rather than aggregate detection accuracy.

### 4.5. Gamma Sensitivity Analysis

To assess the sensitivity of the coupled DSG parameters, $\gamma_c$ and $\gamma_d$ are varied together on the MOT17 reporting split with FRCNN detections. The grid covers $\gamma_c = \gamma_d$ in {0.42,

0.56, 0.70, 0.84, 0.98}. This experiment evaluates the coupled gamma setting only; it does not establish independent sensitivity for every hyperparameter.

The baseline setting $\gamma_c = \gamma_d = 0.70$ is used in the FullModel. The neighboring values in Table 5 provide a local robustness check for this coupled setting, rather than a complete parameter-optimization study.

**Table 5.** Gamma sensitivity of DSG on the MOT17 reporting split with FRCNN detections.

| $\gamma_c$ | $\gamma_d$ | MOTA (%) | IDF1 (%) | IDSw (Total) | Frag (Total) |
|---|---|---|---|---|---|
| 0.42 | 0.42 | 46.42 | 40.00 | 382 | 441 |
| 0.56 | 0.56 | 46.41 | 40.00 | 382 | 442 |
| 0.70 | 0.70 (baseline) | 46.41 | 40.00 | 382 | 442 |
| 0.84 | 0.84 | 46.42 | 40.01 | 381 | 440 |
| 0.98 | 0.98 | 46.43 | 40.08 | 380 | 439 |

The table below reports the gamma grid and the resulting MOTA, IDF1, IDSw, and Frag values.

Across the coupled $\gamma_c/\gamma_d$ grid in {0.42, 0.56, 0.70, 0.84, 0.98}, MOTA varies by 0.02 percentage points, IDF1 by 0.08 points, and total IDSw by two switches (380–382). These results indicate low sensitivity within this tested coupled range.

### *4.6. Cross-Scenario Transfer Analysis on an Industrial Proxy Subset*

To examine the transfer limitations and preliminary transfer potential of the proposed method, experiments are conducted on the VisDrone2019-MOT-val industrial proxy subset. This subset is used to approximate industrial low-altitude conditions, including overhead viewpoints, dense small targets, frequent occlusion, and complex backgrounds. It should be emphasized that this subset is not a complete factory-floor MOT benchmark. Qualitative tracking comparisons for occlusion and homogeneous-appearance scenarios are illustrated in Figure 2, where tracking outputs from OC-SORT and our proposed method are displayed side-by-side across matched video frames. Therefore, the results are used only as proxy evidence and as a way to expose domain-shift failure modes.

As shown in Table 6, the VisDrone proxy table reports MOTA, rather than HOTA. OC-SORT obtains −22.3% MOTA, 7.2% IDF1, 16.4% recall, and 45.2% precision. ByteTrack obtains −21.8% MOTA, 8.1% IDF1, 17.6% recall, and 46.5% precision. The proposed method without additional adaptation obtains −10.5% MOTA and 79.1% precision, but its recall is only 3.6% and its IDF1 is 5.3%. The 3.6% value is recall, not HOTA. The high precision therefore reflects a highly conservative tracker operating on very few positive predictions, not effective industrial MOT. This proxy experiment is retained as a failure-exposure and domain-shift stress test; it cannot support a claim of strong industrial performance.

**Table 6.** Cross-scenario transfer results on the VisDrone2019-MOT-val industrial proxy subset.

| Method | MOTA (%) | IDF1 (%) | IDSw | Precision (%) | Recall (%) |
|---|---|---|---|---|---|
| OC-SORT | −22.3 | 7.2 | 238 | 45.2 | 16.4 |
| ByteTrack | −21.8 | 8.1 | 224 | 46.5 | 17.6 |
| Proposed method | −10.5 | 5.3 | 146 | 79.1 | 3.6 |




**Figure 2.** Qualitative tracking comparison under occlusion and homogeneous appearance in the industrial proxy subset. (**a**) OC-SORT tracking result on a frame with mild partial occlusion; (**b**) OC-SORT tracking result on a frame with severe inter-target occlusion; (**c**) OC-SORT tracking result on a frame with highly homogeneous target appearances; (**d**) OC-SORT tracking result on a frame with dense small targets; (**e**) OC-SORT tracking result on a frame with overhead viewpoint and scale variation; (**f**) OC-SORT tracking result on a frame with crossing target motions; (**g**–**l**) tracking results of the proposed method on the same frames as (**a**–**f**), respectively. Colored bounding boxes with numeric labels denote individual target trajectory identities; overlapping bounding boxes indicate targets in close spatial proximity or under partial occlusion, and are presented to visualize the instance-level association challenge rather than annotation errors.

### *4.7. Runtime Performance Analysis*

Runtime efficiency is evaluated under two protocols. The tracker-level protocol excludes detector inference and is used only for comparison with reproduced baselines; the recorded tracker-level rate is 68.14 FPS. The end-to-end protocol includes YOLOv8s detection, target cropping, FastReID extraction, CLIP inference, FAISS retrieval, and data association, and runs at 21.7 FPS on the RTX 3090 platform. The current logs do not contain a component-level time breakdown, so the separate YOLOv8s, FastReID, CLIP, FAISS, and association shares cannot be reported without rerunning the profiler. The main optimization opportunities are batched crop inference, feature caching, selective CLIP extraction, and asynchronous or lightweight semantic inference.

### *4.8. Failure Analysis in Extremely Dense Scenes*

Although the proposed method improves identity preservation under normal and moderately crowded conditions, failure cases still occur in extremely dense scenes. When

a large number of targets appear in a compact region, severe bounding-box overlap and inaccurate target crops may cause many cropped patches to contain neighboring targets or background regions. This reduces the instance-level discriminability of CLIP visual-semantic features. As a result, the dictionary response distributions of different same-class targets may become highly similar, and multiple candidate detections may satisfy comparable semantic and motion constraints. These problematic conditions are visualised in Figure 3, which highlights baseline identity-switch failures and crop-contamination artefacts caused by heavy bounding-box overlap.

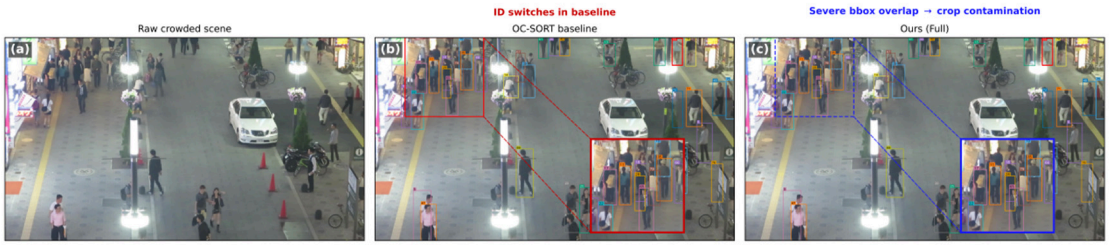


**Figure 3.** Failure-case analysis in extremely dense scenes. (**a**) Overview of an extremely dense crowded scene, showing the raw target distribution and severe bounding-box overlap; (**b**) baseline OC-SORT tracking results on the same scene, where the highlighted overlapping regions indicate identity-switch failures between adjacent same-class targets; (**c**) illustration of crop-contamination artefacts, where the highlighted target crop contains partial visual content of a neighboring target due to heavy bounding-box overlap. Colored bounding boxes with numeric identifiers represent assigned target trajectory identities; overlapping colored boxes in (**b**) mark identity-conflict regions, and the highlighted box in (**c**) indicates the contaminated target crop.

In addition, long-term occlusion weakens the reliability of Kalman prediction. When a target disappears for many frames and reappears in a dense region, multiple candidate detections may satisfy similar motion and semantic constraints. Although DSG can reduce the contribution of unreliable semantic cues, it cannot fully compensate for severe detection misses, inaccurate bounding boxes, or long-term full occlusion. Therefore, the performance boundary of the proposed method is mainly determined by detection quality, crop reliability, and motion prediction stability in ultra-dense scenes.

## 5. Discussion

The experimental results show that category-level CLIP semantics must be reliability constrained when they are used for instance-level online association. In Table 4, fixed-weight CLIP+FAISS produces 558 identity switches and 670 fragmentations, whereas DSG reduces these values to 382 and 442, respectively. Official TrackEval results on the same four-sequence split give HOTA/AssA values of 34.178/26.172 for CLIP+FAISS and 34.542/26.686 for FullModel. Exact reruns with seeds 42, 100, and 2024 produced identical IDSw, MOTA, and IDF1 totals for each configuration; the resulting zero standard deviations indicate deterministic replication, not evidence of stochastic significance. This evidence supports an identity-stability effect; it does not establish a uniform gain in aggregate tracking accuracy.

The proposed dynamic semantic gating mechanism addresses this problem by estimating the reliability of semantic cues at the candidate-pair level. When detection confidence is high, occlusion is weak, motion prediction is consistent, and target density is low, semantic cues can participate more strongly in association. When the association state becomes unreliable, semantic costs are suppressed to avoid misleading matching. This design turns semantic features from fixed association cues into reliability-constrained auxiliary signals.

However, the proposed method has important limitations. First, the MOT17 result is a trade-off: the FullModel reduces identity switches relative to fixed-weight CLIP+FAISS, but its MOTA is 0.84 points lower. Second, the industrial evaluation uses a VisDrone-derived proxy subset, rather than a real factory-floor MOT benchmark; its 3.6% recall and 5.3% IDF1 show that the current system is not effective under this domain shift. Third, HOTA and AssA are reported for seed 42 only, while the exact three-seed reruns cover IDSw, MOTA, and IDF1 and yield zero sample standard deviation. This zero-variance outcome is a deterministic replication result, and does not justify a conventional $p$-value or stochastic-significance claim. Fourth, a complete independent sensitivity analysis for all parameters and a component-level end-to-end latency profile are not documented. Fifth, all MOT17 results use a reporting split from the training set, rather than the official test server, so residual overfitting risk remains. Additional boundary-case failure modes are further elaborated in Appendix A. These limitations define the scope of the current claims and motivate the future experiments below.

## 6. Conclusions

This paper investigates the reliability of CLIP semantic cues in online multi-object tracking and proposes a reliability-aware semantic gating and retrieval-augmented association framework. The main motivation is that CLIP features provide category-level semantic priors, but may introduce instance-level confusion when multiple same-class targets appear densely. To analyze this problem, a semantic discriminative margin, semantic noise rate, and semantic collapse index are introduced, to quantify the semantic ambiguity of CLIP features in trajectory-detection candidate association.

On the MOT17 reporting split, fixed-weight CLIP+FAISS obtains 47.25% MOTA, 39.32% IDF1, 34.178% HOTA, 26.172% AssA, 558 identity switches, and 670 fragmentations, whereas the FullModel obtains 46.41% MOTA, 40.00% IDF1, 34.542% HOTA, 26.686% AssA, 382 identity switches, and 442 fragmentations. DSG therefore reduces identity switches by 31.5% and fragmentations by 34.0%, with a 0.84-point MOTA trade-off and a 0.68-point IDF1 gain. The result supports the narrower claim that DSG improves identity stability under this protocol.

On the MOT17 reporting split with FRCNN detections, the proposed method achieves 46.41% MOTA, 40.00% IDF1, 382 ID switches, and 442 fragmentations. Compared with CLIP+FAISS retrieval, the full model is lower by 0.84 percentage points in MOTA and higher by 0.68 points in IDF1, while also reducing identity switches by 176 and fragmentations by 228, relative to the fixed-weight CLIP+FAISS baseline. The result should therefore be interpreted as a stability gain, rather than a uniform performance win.

Future work will focus on five directions. First, lightweight CLIP inference, feature caching, batch crop processing, and selective semantic extraction will be explored to reduce end-to-end overhead. Second, the prompt templates listed in Appendix B will be evaluated through controlled prompt ablations. Third, broader independent sensitivity studies, confidence intervals where nonzero variance is available, and statistically appropriate tests on independently varying runs will be pursued; the present three-seed result is reported only as deterministic replication. Fourth, official MOT17 test-server evaluation and additional UAV and industrial datasets will be used to reduce training-set evaluation bias. Fifth, a real industrial MOT dataset covering production lines, warehouses, conveyor facilities, and low-altitude inspection scenes will be constructed to evaluate practical deployment.

**Author Contributions:** Conceptualization, R.G. and Y.H.; methodology, R.G.; software, R.G.; validation, R.G. and Y.H.; formal analysis, R.G.; investigation, R.G.; resources, Y.H.; data curation, R.G.; writing—original draft preparation, R.G.; writing—review and editing, Y.H.; visualization, R.G.;

supervision, Y.H.; project administration, Y.H.; funding acquisition, Y.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China (Grant Nos. 11905153, 61701331), the Natural Science Foundation of Sichuan Province (Grant Nos. 2023NSFSC1591, 2022NSFSC0570), the Youth Fund of Humanities and Social Sciences Research of the Ministry of Education of China (Grant No. 23YJC880062), and the Philosophy and Social Science Foundation of Sichuan Province (Grant No. SCJJ24ND156).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The original data presented in the study are openly available in MOTChallenge and VisDrone benchmark websites at https://motchallenge.net/ (accessed on 20 May 2025) and https://visdrone.org/ (accessed on 20 May 2025), respectively. The experimental results reported in this paper are based on these publicly available datasets.

**Acknowledgments:** The authors would like to thank the providers of the MOTChallenge and VisDrone datasets for making their benchmarks publicly available.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| MOT | Multi-Object Tracking |
| CLIP | Contrastive Language-Image Pre-training |
| DSG | Dynamic Semantic Gating |
| FAISS | similarity-search library |
| MOTA | Multi-Object Tracking Accuracy |
| IDF1 | Identity F1 Score |
| IDSw | Identity Switches |
| IoU | Intersection over Union |
| ReID | Re-identification |
| UAV | Unmanned Aerial Vehicle |

## Appendix A. Failure Cases and Limitations

Although the proposed model improves identity stability in several settings, several boundary cases remain. First, in extremely dense scenarios, severe bounding-box overlap and inaccurate target crops may reduce the instance-level discriminability of CLIP visual-semantic features. Second, under long-term full occlusion, Kalman prediction based on the constant-velocity assumption may accumulate errors, making it difficult to recover the original identity when the target reappears. Third, under drastic illumination changes or severe crop misalignment, CLIP visual semantic features may exhibit short-term drift, affecting the stability of semantic association. Future work will improve the method through lightweight semantic encoders, more robust reliability estimation, and validation on real industrial MOT datasets.

## Appendix B. Semantic Dictionary Prompts

The following prompt templates are used by the retrieval-augmented semantic memory. {category} is replaced by the detector class label.

**Table A1.** General category layer.

| Detail Level | Template |
|---|---|
| simple | A photo of a {category} |
| medium | A photo of a {category} in a typical setting |
| detailed | A high-quality photo of a {category} with clear features |

**Table A2.** Industrial domain layer.

| Detail Level | Template |
|---|---|
| simple | A {category} in a warehouse |
| medium | A photo of a {category} in an industrial warehouse setting |
| detailed | A photo of a {category} in an industrial warehouse setting, with typical warehouse lighting and background |

**Table A3.** Warehouse domain layer.

| Detail Level | Template |
|---|---|
| simple | A {category} in a warehouse |
| medium | A photo of a {category} in a warehouse environment |
| detailed | A photo of a {category} in a warehouse environment, with shelves, pallets, and industrial equipment in the background |

The dictionary also uses the fixed three-prompt ensemble: a photo of a {category}, {category} in the scene, and high quality image of {category} when encoding a category name.

## References


1. Dendorfer, P.; Osep, A.; Milan, A.; Schindler, K.; Cremers, D.; Reid, I.; Roth, S.; Leal-Taixé, L. MOTChallenge: A benchmark for single-camera multiple target tracking. *Int. J. Comput. Vis.* **2021**, *129*, 845–881. [CrossRef]
2. Dendorfer, P.; Rezatofighi, H.; Milan, A.; Shi, J.; Cremers, D.; Reid, I.; Roth, S.; Schindler, K.; Leal-Taixé, L. MOT20: A benchmark for multi object tracking in crowded scenes. *arXiv* **2020**, arXiv:2003.09003.
3. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing*; IEEE: New York, NY, USA, 2016; pp. 3464–3468.
4. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing*; IEEE: New York, NY, USA, 2017; pp. 3645–3649.
5. Cao, J.; Pang, J.; Weng, X.; Khirodkar, R.; Kitani, K. Observation-centric SORT: Rethinking SORT for robust multi-object tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 9686–9696.
6. Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; Wang, X. ByteTrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022; pp. 1–21.
7. Du, Y.; Zhao, Z.; Song, Y.; Zhao, Y.; Su, F.; Gong, T.; Meng, H. StrongSORT: Make DeepSORT great again. *IEEE Trans. Multimed.* **2023**, *25*, 8725–8737. [CrossRef]
8. Maggiolino, G.; Ahmad, A.; Cao, J.; Kitani, K. Deep OC-SORT: Multi-pedestrian tracking by adaptive re-identification. In *2023 IEEE International Conference on Image Processing*; IEEE: New York, NY, USA, 2023; pp. 3025–3029.
9. Jocher, G.; Chaurasia, A.; Qiu, J. YOLO by Ultralytics [EB/OL]. 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 20 May 2025).
10. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
11. Diwan, T.; Anirudh, G.; Tembhurne, J.V. Object detection using YOLO: Challenges, architectural successors, datasets and applications. *Multimed. Tools Appl.* **2023**, *82*, 9243–9275. [CrossRef] [PubMed]

12. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S.C.H. Deep learning for person re-identification: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 2872–2893. [CrossRef] [PubMed]
13. He, L.; Liao, X.; Liu, W.; Liu, X.; Cheng, P.; Mei, T. FastReID: A PyTorch toolbox for general instance re-identification. *arXiv* **2020**, arXiv:2006.02631.
14. Miao, J.; Wu, Y.; Liu, P.; Ding, Y.; Yang, Y. Pose-guided feature alignment for occluded person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 542–551.
15. Sun, Y.; Cheng, C.; Zhang, Y.; Zhang, C.; Zheng, L.; Wang, Z.; Wei, Y. Circle loss: A unified perspective of pair similarity optimization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6398–6407.
16. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2021; pp. 8748–8763.
17. Li, S.; Fischer, T.; Ke, L.; Ding, H.; Danelljan, M.; Yu, F. OVTrack: Open-vocabulary multiple object tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 5567–5576.
18. Wu, J.; Jiang, Y.; Liu, Q.; Yuan, Z.; Bai, X.; Bai, S. General Object Foundation Model for Images and Videos at Scale. *arXiv* **2023**, arXiv:2312.09158. [CrossRef]
19. Zhu, P.; Wen, L.; Du, D.; Bian, X.; Fan, H.; Hu, Q.; Ling, H. Detection and tracking meet drones challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7380–7399. [CrossRef] [PubMed]
20. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object detection in 20 years: A survey. *Proc. IEEE* **2023**, *111*, 257–276. [CrossRef]
21. Lv, W.; Xu, S.; Zhao, Y.; Wang, G.; Wei, J.; Cui, C.; Du, Y.; Dang, Q.; Liu, Y. DETRs beat YOLOs on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 16–22 June 2024.
22. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 87–110. [CrossRef] [PubMed]
23. Aharon, N.; Orfaig, R.; Bobrovsky, B.Z. BoT-SORT: Robust associations multi-pedestrian tracking. *arXiv* **2022**, arXiv:2206.14651.
24. Zhang, Y.; Wang, C.; Wang, X.; Zeng, W.; Liu, W. FairMOT: On the fairness of detection and re-identification in multiple object tracking. *Int. J. Comput. Vis.* **2021**, *129*, 3069–3087. [CrossRef]
25. Bergmann, P.; Meinhardt, T.; Leal-Taixe, L. Tracking without bells and whistles. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 941–951.
26. Zhou, X.; Koltun, V.; Krahenbuhl, P. Tracking objects as points. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 474–490.
27. Meinhardt, T.; Kirillov, A.; Leal-Taixe, L.; Feichtenhofer, C. TrackFormer: Multi-object tracking with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 8844–8854.
28. Zeng, F.; Dong, B.; Zhang, Y.; Wang, T.; Zhang, X.; Wei, Y. MOTR: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022; pp. 659–675.
29. Zhang, Y.; Wang, T.; Zhang, X. MOTRv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 22056–22065.
30. Sun, P.; Cao, J.; Jiang, Y.; Yuan, Z.; Bai, S.; Kitani, K.; Luo, P. DanceTrack: Multi-object tracking in uniform appearance and diverse motion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 20993–21002.
31. Cui, Y.; Zeng, C.; Zhao, X.; Yang, Y.; Wu, G.; Wang, L. SportsMOT: A large multi-object tracking dataset in multiple sports scenes. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 9921–9931.
32. Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; Jitsev, J. Reproducible scaling laws for contrastive language-image learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 2818–2829.
33. Zhai, X.; Mustafa, B.; Kolesnikov, A.; Beyer, L. Sigmoid loss for language image pre-training. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 11975–11986.
34. Gu, X.; Lin, T.Y.; Kuo, W.; Cui, Y. Open-vocabulary object detection via vision and language knowledge distillation. In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022.
35. Li, L.H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. Grounded language-image pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 10965–10975.

36. Zhong, Y.; Yang, J.; Zhang, P.; Li, C.; Codella, N.; Li, L.H.; Zhou, L.; Dai, X.; Yuan, L.; Li, Y.; et al. RegionCLIP: Region-based language-image pretraining. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 16793–16803.
37. Zhou, X.; Girdhar, R.; Joulin, A.; Krahenbuhl, P.; Misra, I. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022; pp. 350–368.
38. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set detection. *arXiv* **2023**, arXiv:2303.05499.
39. Li, J.; Li, D.; Savarese, S.; Hoi, S. BLIP-2: Bootstrapping language-image pre-training with frozen encoders and large language models. In *International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2023; pp. 19730–19742.
40. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y.; et al. Segment Anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 4015–4026.
41. Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; Tian, Q. The Unmanned Aerial Vehicle Benchmark: Object Detection and Tracking. In *Computer Vision—ECCV 2018 Workshops*; Springer: Cham, Switzerland, 2019; pp. 370–386.
42. Wen, L.; Du, D.; Cai, Z.; Lei, Z.; Chang, M.C.; Qi, H.; Lim, J.; Yang, M.H.; Lyu, S. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. *Comput. Vis. Image Underst.* **2020**, *193*, 102907. [CrossRef]
43. Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; Darrell, T. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2636–2645.
44. Kalman, R.E. A new approach to linear filtering and prediction problems. *J. Basic Eng.* **1960**, *82*, 35–45. [CrossRef]
45. Kuhn, H.W. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **1955**, *2*, 83–97. [CrossRef]
46. Luiten, J.; Osep, A.; Dendorfer, P.; Torr, P.; Geiger, A.; Leal-Taixé, L.; Leibe, B. HOTA: A higher order metric for evaluating multi-object tracking. *Int. J. Comput. Vis.* **2021**, *129*, 548–578. [CrossRef] [PubMed]
47. Johnson, J.; Douze, M.; Jegou, H. Billion-scale similarity search with GPUs. *IEEE Trans. Big Data* **2021**, *7*, 535–547. [CrossRef]


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.