

Appearance-Guided Vehicle Tracking and Speed Estimation for Low-Frame-Rate Industrial CCTV

Jaehun Hwang¹, Keonwoo Park¹, Gyulim Gu², Jinhong Min², Bongsu Cho³, Minseok Song^{1,†}

¹ Department of Industrial and Management Engineering
Pohang University of Science and Technology (POSTECH), Pohang, Republic of Korea
Emails: jaehunh@postech.ac.kr; pkwoo2001@postech.ac.kr; mssong@postech.ac.kr

² LG Electronics; ³ spectrabrain
Emails: gyulim.gu@lge.com; jhong.min@lge.com; bongsu.cho@spectrabrain.ai

Abstract. Industrial CCTV analytics must associate vehicles across sparse observations before estimating physical speed. We present AppearanceTracking, which fuses pretrained deep features, HSV histograms, and HOG descriptors with guarded multi-object association. Iterative proposals use fixed competition for approval, singleton candidate sets use raw similarity thresholds, and image motion provides a bounded penalty. User-specified road dimensions define a homography for timestamp-based speed estimation. On 12 UA-DETRAC sequences at 1, 2, 5, 10, and 25 FPS, shared detections and fixed tracker-specific configurations enable comparison with five established trackers. AppearanceTracking leads HOTA, IDF1, and MOTA at 1–2 FPS. At 1 FPS, its 57.27% HOTA and 68.83% IDF1 exceed the best baseline by 11.18 and 24.42 percentage points. On nine BrnoCompSpeed views, offline benchmark scoring at 10 FPS yields 1.08 km/h mean absolute error (MAE) with 92.3% coverage under reference calibration. A separate 1–10 FPS comparison with common relaxed observation requirements and a time-based pair-median estimator yields 1.26 km/h MAE and 82.4% coverage at 5 FPS using half the 10-FPS inputs, while coverage falls sharply at 1–2 FPS. These results connect sparse-frame association to camera-specific sampling choices based on speed accuracy and coverage.

Keywords: low-frame-rate CCTV; multi-object tracking; appearance fusion; homography; speed estimation

1. INTRODUCTION

Vehicle speed requires identity association, metric scene geometry, and elapsed time. A camera supplies only image observations; errors in any stage can propagate into the resulting physical speed.

Multi-camera industrial monitoring must accommodate many CCTV streams within a shared processing infrastructure. At a fixed analysis FPS, adding cameras increases the recurring work of detection, feature extraction, and tracking, potentially requiring additional GPUs or distribution across servers. Maintaining tracking quality at a lower analysis FPS offers a way to accommodate more streams within a fixed inference budget, rather than expanding infrastructure solely to sustain high-rate analysis. We therefore view low-FPS tracking not only as per-stream computation reduction, but as an approach to integrated operation of industrial CCTV under limited infrastructure.

Collecting representative training data can also be costly, and transfer to new views or operating conditions is not guaranteed. We therefore target fixed monocular cameras at 1–2 frames per second (FPS), using pretrained features without additional encoder or association-network training.

Figure 1 illustrates larger image displacements under sparse sampling, which can weaken motion prediction and

overlap-based association. Industrial scenes also involve turning, reversing, stationary vehicles, and partial occlusion, limiting reliance on motion continuity alone. Appearance supplies another cue, but partial boxes, similar vehicles, and stale tracks still compete. Wrong matches can contaminate stored appearance and subsequent decisions, motivating explicit approval and state maintenance.

AppearanceTracking supplies IDs and image-position histories; measured road geometry and capture timestamps support speed estimation. Association requires no calibration, and physical speed does not feed back into matching.

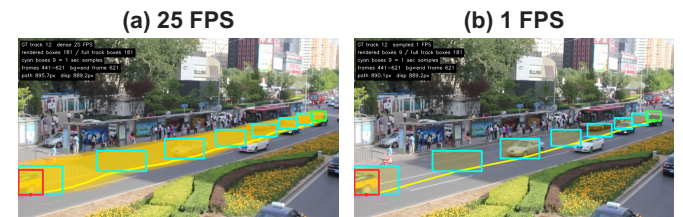


Figure 1: One UA-DETRAC GT vehicle at 25 FPS (a) and 1 FPS (b), overlaid on the final frame (Wen et al., 2020). Boxes are observations, not predictions.

Our contributions are: (1) appearance-guided association with directional fusion, iterative approval, singleton handling, and bounded state; (2) measured-rectangle calibration and timestamp-based speed integration; and (3) six-tracker comparisons and controlled ablations at five sampling rates on 12

† Corresponding author: Minseok Song, mssong@postech.ac.kr

SCENE SETUP | independently for each camera

Four ordered image points on the road
Measured width W and length L
Physical rectangle, not an arbitrary image box

Compute homography H
Pixels $\rightarrow$ road-plane meters

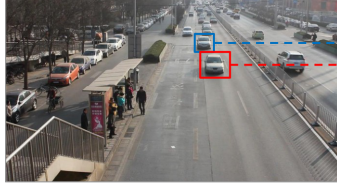
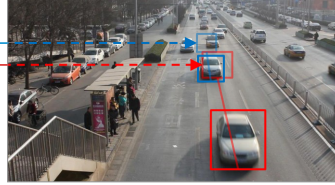
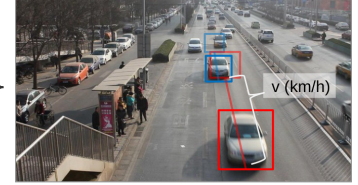
ONLINE PROCESSING | sampled frames and capture times**Vehicle detection****AppearanceTracking****Projection + speed**

Figure 2: Photo-based overview with illustrative overlays on UA-DETRAC frames. Each camera has independent calibration and tracking state. H supplies the metric mapping for speed estimation; association uses image-space observations.

UA-DETRAC sequences, complemented by calibration analyses and FPS-dependent speed-accuracy and coverage evaluation on nine BrnoCompSpeed views.

2. RELATED WORK

SORT combines Kalman filtering and Hungarian assignment (Bewley et al., 2016), and Deep SORT adds a learned appearance metric (Wojke et al., 2017). ByteTrack recovers tracklets using low-score detections (Zhang et al., 2022). BoT-SORT combines motion, appearance, and camera-motion compensation (Aharon et al., 2022), while OC-SORT revisits observations to improve motion estimation during occlusion (Cao et al., 2023).

MOTR learns end-to-end temporal association using track queries (Zeng et al., 2022), while MOTIP formulates association as trainable in-context ID prediction (Gao et al., 2025). Both learn from annotated sequences and report strong benchmark performance.

Our approach instead uses explicit association rules with fixed pretrained features. Appearance is primary, while timed image-motion history provides a bounded complement. The focus is proposal acceptance and state management, including candidate sets with uninformative relative normalization.

DINOV2 (Oquab et al., 2024), color histograms, and HOG (Dalal and Triggs, 2005) provide complementary appearance cues. Their independently computed scores are fused after directional normalization, without encoder fine-tuning.

For speed measurement, image motion must be related to a metric road plane. Planar projective geometry supplies this mapping when suitable correspondences are known (Hartley and Zisserman, 2004). BrnoCompSpeed provides independent speed references and calibration metadata for studying this problem (Sochor et al., 2018). We use an operator-defined physical rectangle to avoid requiring automatic extraction of vanishing-point evidence from every installation.

3. FRAMEWORK

3.1 Problem Setting and State

Let n index analyzed sampled frames and τ_n their capture times (seconds). Frames analyzed without detections increment n ; unsampled source frames do not. At each step, N detections are compared with M non-deleted tentative, active, or lost tracks. Each camera-local track stores its ID, lifecycle state, latest observed box, appearance descriptors, and bounded motion/ground-position histories.

Figure 2 proceeds from detection and features to association, state maintenance, and speed. Empty candidate sets bypass assignment. H is fixed per sequence; calibration and metric speed do not influence association. Motion uses box centers; speed uses observed bottom-center anchors.

3.2 Detection and Appearance Representation

Vehicle detections are obtained with the pretrained RT-DETR R50 model PekingU/rt detr_r50vd_coco_o365 (Zhao et al., 2024). Boxes are clipped to image bounds and checked for validity and size before feature extraction. Confidence filtering and duplicate removal are specified for the tracking benchmark in Section 5.1.

Each crop yields a 768-dimensional DINOv2-base CLS descriptor, a 32-dimensional HSV descriptor, and a 1,764-dimensional HOG descriptor. DINOv2 uses RGB input, shorter-side resizing to 256, a 224×224 center crop, 1/255 rescaling, ImageNet normalization, and FP32 inference. HSV concatenates independently L1-normalized H/S/V histograms with 16/8/8 bins. HOG uses 64×64 grayscale crops, nine unsigned orientation bins, 8×8-pixel cells, and 2×2-cell L2-Hys blocks with one-cell stride. No encoder is fine-tuned.

These descriptors respond differently to crop content. HSV counts colors without encoding their spatial arrangement, whereas DINOv2 summarizes learned visual patterns and

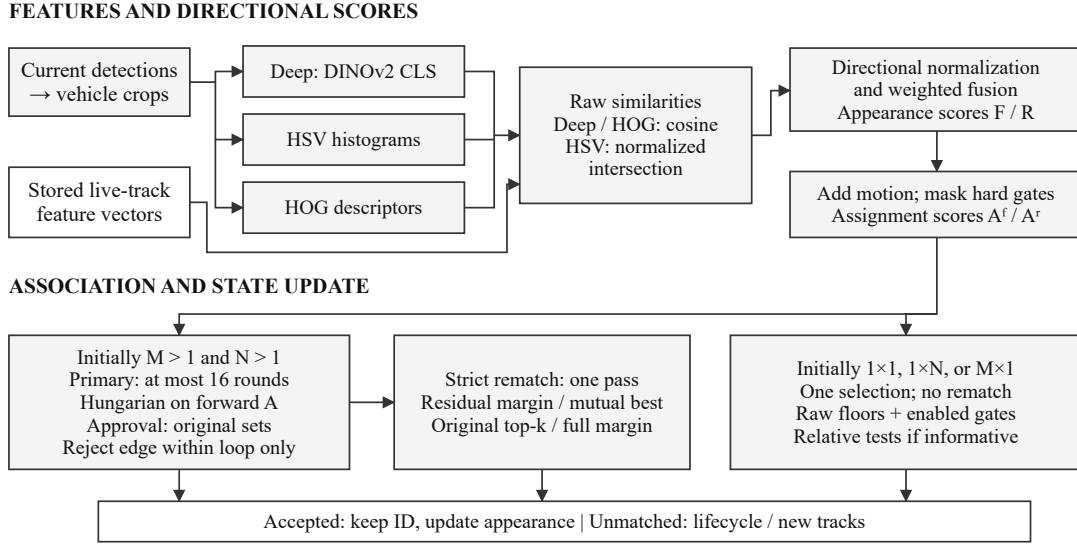


Figure 3: AppearanceTracking separates appearance floors from adjusted-score assignment and validation. Branching uses the initial candidate dimensions; primary approval retains the original competition.

HOG retains local edge structure. Bounding-box crops can include background or occluding vehicles: these pixels directly change HSV counts and may also alter the deep representation. Neither cue is assumed to be foreground-only.

3.3 Directional Fusion and Match Approval

For track i and detection j , s_{ij}^k is the raw similarity for component k . Deep and HOG use cosine similarity. For HSV vectors h_i and h_j , normalized intersection is $\Sigma_{\beta} \min(h_{i\beta}, h_{j\beta}) / \min(\Sigma_{\beta} h_{i\beta}, \Sigma_{\beta} h_{j\beta})$. For a candidate-score vector s , let $m = \text{median}(s)$ and $d = 1.4826 \text{ median}|s - m|$. Normalize using

$$\mathcal{N}(s)_j = \frac{s_j - m}{d}. \quad (1)$$

A robust scale $d \leq 10^{-12}$ triggers ordinary z-score normalization with population standard deviation; if that also vanishes, all scores become zero. Even-length medians average the two middle values. Forward and reverse appearance scores are

$$F_{ij} = \sum_k w_k \mathcal{N}(s_{i,:}^k)_j, \quad R_{ij} = \sum_k w_k \mathcal{N}(s_{:,j}^k)_i. \quad (2)$$

Both matrices use track–detection indexing. Normalization is over the original row or column, with Deep/HSV/HOG weights 0.3/0.5/0.2. F and R can differ despite symmetric raw similarities; they are relative scores, not probabilities.

The color-emphasized configuration reflects development-stage observations: HSV was useful in daylight scenes with relatively uniform illumination and when the visible side of a vehicle changed. Its distributional representation complements shape-based cues when sparse sampling produces larger position or pose changes. The retained Deep and HOG terms provide learned appearance and local edge information, respectively, while shadows, glare, and low illumination can

weaken color evidence. These weights remain a fixed engineering setting informed by development observations. Section 5.2 evaluates the conditional value of including HSV and HOG, rather than the optimality of this particular ratio.

Before hard-gate masking, assignment scores are

$$A_{ij}^F = F_{ij} + P_{ij}, \quad A_{ij}^R = R_{ij} + P_{ij}. \quad (3)$$

Here $P_{ij} \leq 0$ is the image-motion penalty (Section 3.5). Pairs outside the area-ratio range 0.125–8, measured against the last observed box, are then masked in A. Thus, gated pairs still influence the preceding normalization. Hungarian assignment maximizes forward A^f only, not a combination of directions. Approval uses F/R for score floors and masked A^f/A^r for margins and mutual-best checks.

The forward margin is the selected adjusted score minus its largest competitor in the specified set. In regular primary matching and rematching, competitors with $\text{IoU} \geq 0.9$ relative to the selected detection are excluded from this margin only, not from normalization, mutual-best checks, or singleton margins. Mutual best requires detection j to be track i 's forward best and track i to be detection j 's reverse best. An undefined required margin causes rejection; reverse-margin thresholds are disabled.

Primary approval requires $F, R \geq 0.5$, a forward adjusted margin ≥ 0.2 , and mutual best in the original candidate sets. Iterative primary matching changes only the assignment subproblem: accepted endpoints are removed and rejected edges are forbidden within the primary loop. Normalization, approval margins, and mutual-best reference sets remain fixed at their original full-frame values. At most 16 rounds are followed by one strict rematch pass. Assignment alone never authorizes an ID update.

3.4 Residual and Singleton Candidates

Strict rematch revisits unresolved candidates without renormalizing scores: $F, R \geq 1.0$, a residual forward adjusted margin ≥ 0.5 and residual mutual best, plus original-set bilateral top-3 eligibility in masked assignment scores and full-row adjusted margin ≥ 0.2 are required. Removing accepted competitors permits recovery under stricter thresholds. If either initial candidate dimension is one, only the singleton branch runs, requiring raw floors $\text{Deep} \geq 0.53$, $\text{HSV} \geq 0.65$, $\text{HOG} \geq 0.50$, sampled-frame gap one, and enabled hard gates. A 1×1 pair needs no relative tests; for $1 \times N$ or $M \times 1$, the informative direction selects the highest adjusted-score pair and additionally requires appearance score ≥ 0.5 and adjusted margin ≥ 0.2 . Rejection triggers neither a runner-up nor rematching.

3.5 Motion, Lifecycle, and Memory

Image motion uses the latest three timed box centers. Adjacent displacements divided by capture-time differences give velocities; their component-wise median predicts the current center from the last observation. The normalized error d_{ij} is the center prediction distance divided by the mean diagonal length of the previous and candidate boxes, lower-bounded by one pixel. The bounded penalty is

$$P_{ij} = -\lambda \rho_i \left[1 - \exp \left(-\frac{d_{ij}^2}{2\sigma^2 U_i^2} \right) \right]. \quad (4)$$

Here $\lambda=4$ and $\sigma=1$. For a valid prediction, let δ be the time since the last observation, h the median valid adjacent interval, K the number of valid velocity intervals, and b the last box diagonal (at least one pixel). With adjacent velocities u_r and their component-wise median $\bar{u}$, define $D = \text{median}_r \|u_r - \bar{u}\|_2$ and $C_v = \max\{\|\bar{u}\|_2, 0.1b/h, \varepsilon\}$, where $\varepsilon=10^{-12}$ is the numerical floor. A prediction requires at least two valid adjacent intervals; each interval and δ must exceed ε . The gap-dependent scale is $U=1+0.25 \max(0, \delta/h-1)$, and confidence is $\rho = \text{clip}(\min(1, K/2) \exp(-D/C_v)/U, 0, 1)$; track subscripts are omitted here. Confidence thus decreases with inconsistent velocities and prolonged gaps. With three observations and two valid intervals, the support factor is one. If prediction is unavailable or the track is tentative, $P_{ij}=0$. Three prior observations are required, so this prior normally starts at the fourth observation. The separate distance hard gate is disabled.

Scene-exit deletion applies only to unmatched active tracks. Four observations support a robust velocity for consistency checking, while the latest two observations provide the velocity used to predict an image-boundary crossing. Reliable outward motion, a last box within 16 pixels of the boundary, and a sufficiently short prediction gap are required; matched observations take precedence.

Unmatched detections create tentative tracks with IDs and an initial hit count of one. Two consecutive detection-supported steps confirm a track; a tentative miss resets this count.

On an unmatched step, tentative tracks are deleted when miss count exceeds two or age exceeds five sampled frames. Confirmed tracks become lost on the first miss and are deleted after more than three consecutive misses, unless scene-exit deletion applies. Recovery preserves the ID. Accepted associations replace appearance descriptors with the latest observation; this is separate from speed EMA.

4. CALIBRATION AND SPEED

4.1 User-Assisted Metric Calibration

The user-assisted design makes calibration an explicit, editable setup step for each fixed camera. An operator can choose a measurable rectangle on the road plane of interest and supply its physical dimensions, rather than requiring the system to infer vanishing points from lane structure or straight vehicle trajectories. The same interaction applies across installations while leaving the reference geometry and scale under operator control.

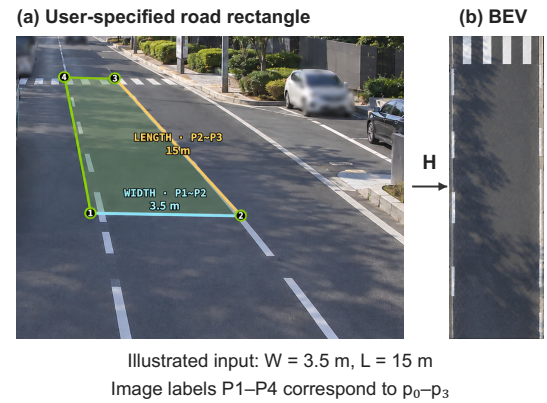


Figure 4: User-defined road rectangle and its metric bird's-eye view. P1–P4 correspond to p_0 – p_3 .

The operator supplies four ordered image points $p_0, \dots, p_3$ defining calibration quadrilateral C , the image of a physical road rectangle. Edges $p_0 \rightarrow p_1$ and $p_1 \rightarrow p_2$ have measured lengths W and L , defining metric targets $(0,0)$, $(W,0)$, (W,L) , and $(0,L)$ (Figure 4). C need not be rectangular in the image. H maps homogeneous image coordinates to metric coordinates:

$$\tilde{q} \sim H\tilde{p}, \quad q = (\tilde{q}_1/\tilde{q}_3, \tilde{q}_2/\tilde{q}_3). \quad (5)$$

H is computed directly from the four image-to-metric correspondences (Hartley and Zisserman, 2004). Inputs must be finite and non-degenerate, W, L positive, and H numerically stable. The formulation assumes measured rectangle dimensions on a fixed road plane; radial distortion is not estimated.

Once configured, H is reused while the camera view and road plane remain unchanged. The operator can correct the selected points or measured dimensions and recompute H without modifying the tracking algorithm. This trades an initial measurement input for a calibration basis that can be directly inspected and revised.

4.2 Timestamp-Based Speed Estimation

After association, an observed box (x_1, y_1, x_2, y_2) supplies the bottom-center anchor $p_n = ((x_1 + x_2)/2, y_2)$. C defines H without imposing a spatial gate, so anchors across the image are projected under the same road-plane assumption. Speed estimation begins when the second consecutive detection-supported hit confirms a track, using the initial matched pair subject to Section 4.3.

The history retains only finite observed ground positions, not predictions: at most five from the latest two seconds, with a indexing the oldest after pruning. The sample cap can shorten this window at higher sampling rates (Figure 5):

$$v_n^{\text{raw}} = 3.6 \frac{\|q_n - q_a\|_2}{\tau_n - \tau_a} \quad (6)$$

Elapsed time uses capture timestamps in seconds. This is finite-window endpoint displacement, not instantaneous velocity or the sum of intermediate path lengths; curved motion can be underestimated.

4.3 Validity Checks and Smoothing

At least two projected samples and $\Delta\tau \geq 0.02$ s are required; otherwise, no estimate is produced. Displacement below 0.15 m gives measured zero and resets the EMA to zero. Non-finite speed estimates and those exceeding 250 km/h are rejected. Outside the deadband, an accepted raw speed initializes the EMA if no prior smoothed estimate exists; otherwise,

$$v_n = 0.5v_n^{\text{raw}} + 0.5v_{\text{prev}} \quad (7)$$

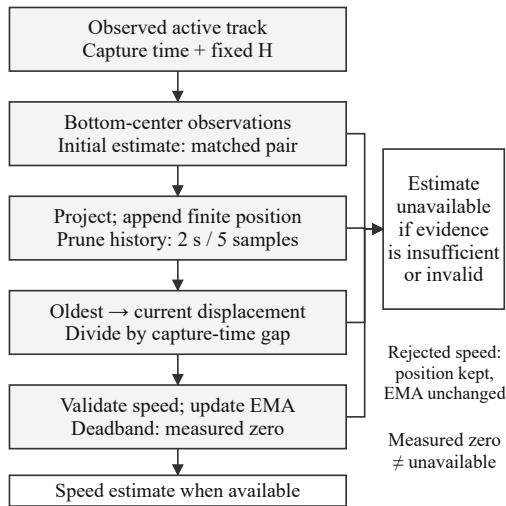


Figure 5: Active-track speed across the scene; the initial matched pair starts estimation.

Here v_{prev} is the preceding smoothed estimate. A rejected speed leaves the EMA unchanged while retaining its projected position. Projection failure adds no position, and expiry of all previous samples restarts estimation.

5. EVALUATION

5.1 Tracking Performance Across Sampling Rates

Protocol. Twelve UA-DETRAC sequences (Wen et al., 2020) are sampled at 1, 2, 5, 10, and 25 FPS. Nearest-frame sampling from 25-FPS sources preserves actual capture timestamps (maximum timing error 20 ms); inputs and GT use the same sampled frames.

A shared RT-DETR R50vd COCO/O365 bank, generated at confidence 0.1 after near-duplicate and contained-part removal, supplies all trackers. Input cutoffs are 0.1 for ByteTrack/BoT-SORT and 0.5 otherwise. Detections and features are reused at every FPS. DeepSORT/BoT-SORT use FastReID; ours uses DINOv2-base, HSV, and HOG. Parameters remain fixed across FPS; native ByteTrack/BoT-SORT buffer scaling is retained, while other frame-based lifetimes remain unchanged.

TrackEval pools HOTA, IDF1, and MOTA across sequences (Luiten et al., 2021). Vehicle annotations and ignore regions are retained, with CLEAR/Identity IoU 0.5, ignore IoA 0.5, and the default HOTA IoU sweep. This is a fixed-detector comparison rather than the official UA-DETRAC PR-MOTA protocol.

Results. AppearanceTracking leads HOTA, IDF1, and MOTA at 1–2 FPS (Table 1). At 1 FPS, it exceeds the best baseline by 11.18, 24.42, and 17.23 percentage points, respectively.

At 5 FPS, AppearanceTracking leads HOTA and MOTA, whereas BoT-SORT leads IDF1. BoT-SORT leads all three metrics at 10–25 FPS. From 25 to 1 FPS, it loses 7.09 HOTA points compared with 21.30–30.15 for the baselines (Figure 6). The advantage is thus specific to sparse observations rather than uniform across sampling rates.

5.2 Tracking Ablation

Protocol. To distinguish feature evidence from association design, we compare seven configurations on the same 12 UA-DETRAC segments at 1, 2, 5, 10, and 25 FPS (Section 5.1). Detections (cutoff 0.5), cached features, timestamps, and TrackEval settings are shared, while tracker states are independent. Parameters and lifecycle/output policies remain fixed except for the changes below, without variant-specific retuning.

Feature configurations. Deep-only, HSV-only, Deep+HSV, and Full retain the complete association structure. Their Deep/HSV/HOG weights are 1/0/0, 0/1/0, 0.375/0.625/0, and 0.3/0.5/0.2, respectively. Deep+HSV preserves the Deep:HSV ratio of 3:5 and renormalizes it to sum to one. Excluded features affect neither fusion nor singleton approval; retained features use the raw floors in Section 3.4.

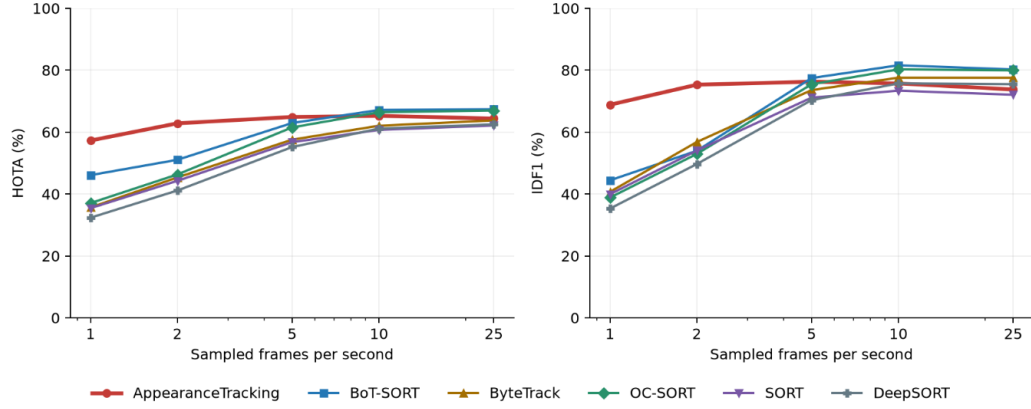


Figure 6: Pooled HOTA and IDF1 across sampling rates (logarithmic FPS axes). All six tracker use fixed configurations.

Table 1: Low-FPS tracking results pooled over 12 UA-DETRAC sequences (%). Best in each column is bold.

Tracker	1 FPS			2 FPS			5 FPS		
	HOTA ↑	IDF1 ↑	MOTA ↑	HOTA ↑	IDF1 ↑	MOTA ↑	HOTA ↑	IDF1 ↑	MOTA ↑
SORT	35.42	39.92	22.98	44.25	54.03	37.91	56.73	71.13	61.81
DeepSORT	32.34	35.28	16.11	41.12	49.67	29.13	55.16	70.31	54.43
ByteTrack	35.65	40.68	24.35	45.34	56.79	36.14	57.55	73.57	55.61
BoT-SORT	46.09	44.41	36.19	51.04	53.95	43.34	62.93	77.45	66.05
OC-SORT	37.06	38.83	22.11	46.31	52.93	34.61	61.52	75.44	61.95
AppearanceTracking	57.27	68.83	53.42	62.83	75.30	63.91	64.88	76.28	69.14

Table 2: Pooled ablations at 1–2 FPS (%; best in bold).

Point differences use unrounded scores.

FPS	Configuration	HOTA ↑	IDF1 ↑	MOTA ↑
1	Full method	57.27	68.83	53.42
	Deep-only	49.41	57.67	44.92
	HSV-only	52.38	61.54	48.40
	Deep+HSV	55.01	65.19	51.43
	One-pass	53.82	62.84	53.15
	Primary 1-round	54.54	65.39	50.27
	No motion penalty	54.43	64.26	52.45
	Full method	62.83	75.30	63.91
2	Deep-only	58.02	67.95	59.20
	HSV-only	60.41	71.44	61.78
	Deep+HSV	62.51	74.60	63.65
	One-pass	59.06	69.21	62.39
	Primary 1-round	60.20	71.92	61.34
	No motion penalty	59.22	69.28	61.91

Association variants. One-pass uses one regular primary Hungarian round without primary margin/mutual-best checks or strict rematch, retaining score floors, hard gates, motion, and singleton handling. Primary 1-round limits only primary rounds, retaining all approval rules, strict rematch, and singleton handling. No motion penalty sets only the association penalty weight to zero; motion history and scene-exit deletion remain enabled. Table 2 reports the pooled results at 1–2 FPS; higher-rate findings are summarized below.

Feature contributions. Deep+HSV improves pooled HOTA, IDF1, and MOTA over Deep-only at every tested rate, with larger gains under sparser sampling. The respective gains are 5.61/7.52/6.51 points at 1 FPS, compared with 1.32/2.97/0.44 at 25 FPS. Deep+HSV also outperforms HSV-only at every rate, supporting complementary color and deep evidence. Adding HOG provides its clearest benefit at 1 FPS, but at 5 FPS Deep+HSV exceeds Full by 0.06/0.20 HOTA/IDF1 points while Full retains higher MOTA.

Association design. Using the same fused features, Full outperforms One-pass, supporting the joint value of repeated proposals and guarded approval/recovery. At 1 FPS, the modest MOTA gain reflects fewer ID switches and false positives being partly offset by more false negatives. Limiting primary matching to one round also reduces HOTA and IDF1 despite retaining approval rules and strict rematch, showing that repeated proposals add value beyond the recovery pass. Removing only the motion penalty reduces performance, particularly at low FPS, supporting bounded motion as a complement to appearance under sparse observations.

5.3 Speed Estimation: Calibration and Sampling

We evaluate BrnoCompSpeed sessions 4–6 using their center, left, and right views (nine views; Sochor et al., 2018). For the calibration comparison, AppearanceTracking generates bottom-center trajectories at 10 FPS, and the benchmark median estimator provides one representative speed per track. Released trajectories from Dubská et al. (2014) provide an additional, algorithm-generated source for this comparison.

Table 3: Session-level mean absolute error (km/h) by rectangle (A: reference; B: short; C: long) and trajectory source. Each value is the equally weighted mean of the center, left, and right view MAEs.

Session	Reference	Reference	Reference	Proposed	Proposed	Proposed
	A	B	C	A	B	C
S4	1.28	9.86	1.81	0.78	13.59	1.26
S5	2.34	7.22	2.56	0.90	6.00	1.83
S6	1.24	2.61	2.19	1.54	3.09	1.91

Tracks are linked to reference vehicles using lane and measurement-line crossing time; each matched reference contributes one absolute error. Coverage is the fraction of eligible reference records with both a representative speed and a reference correspondence; the AppearanceTracking runs share 2,120 eligible records. MAE is averaged equally across views and describes the matched vehicles.

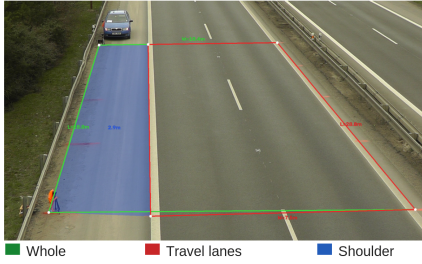


Figure 7: Calibration-region placement example (S4 center view): whole area, travel lanes, and shoulder.

A is the dataset’s measured road rectangle. Its corners and physical width W and length L define H_A . We map the corners of the short B and long C rectangles through H_A to obtain road-plane coordinates and derive their dimensions, then compute a homography for each rectangle. This uses A’s metric mapping for road and shoulder regions on the same plane. Trajectories and matched reference vehicles remain fixed across calibration conditions. Figure 7 illustrates a separate placement comparison.

In Table 3, C yields lower session-averaged MAE than B in S4–S6 for both trajectory sources, although the S5 left and S6 left views are exceptions at the view level. With the proposed 10-FPS trajectories, B reaches 32.97 km/h in the S4 right view, and the nine-view mean MAEs for A, B, and C are 1.08, 7.56, and 1.67 km/h, respectively, with 92.3% coverage. These differences on fixed trajectories highlight the importance of calibration-region selection and metric dimensions; rectangle size and placement vary jointly.

A separate 10-FPS test in the S4 center view varied only the entered length while holding each road-only rectangle’s corners, width (7.5 m), and proposed trajectories fixed. For reference lengths near 7, 14, and 28 m, a +1 m input error produced relative speed errors (absolute error divided by reference speed) of 14.6%, 7.4%, and 3.8%, respectively, showing reduced sensitivity with longer references.

For the sampling-rate comparison (Table 4), AppearanceTracking is run independently at 1, 2, 5, and 10 FPS over the same ten-minute intervals, with calibration A fixed within each view. Both offline scoring settings are held constant across rates: Relaxed requires at least three observations overall and two between measurement lines with a time-based pair-median estimator; Benchmark requires six in both cases with benchmark median scoring.

Table 4: Vehicle-level speed MAE and coverage across sampling rates with fixed measured-road calibration.

FPS	Matched n	Coverage (%)	MAE (km/h)
1	74 / 0	3.5 / 0.0	16.74 / —
2	866 / 0	40.8 / 0.0	2.26 / —
5	1,747 / 1,380	82.4 / 65.1	1.26 / 1.08
10	1,960 / 1,957	92.5 / 92.3	1.20 / 1.08

Values: Relaxed / Benchmark; —: unavailable MAE (no matches).

Under Relaxed, 5 FPS uses half the 10-FPS input frames while retaining 82.4% coverage and 1.26 km/h MAE, versus 92.5% and 1.20 km/h at 10 FPS (Table 4). At 1–2 FPS, coverage falls sharply, with a marked error increase at 1 FPS. Under Benchmark, 5 and 10 FPS have the same reported MAE despite substantially different coverage; no vehicles are evaluable at 1–2 FPS. MAE describes the vehicles evaluable at each rate; sampling should therefore reflect both accuracy and required measurement coverage.

6. DISCUSSION

Sampling rate should be treated as a resource-allocation choice within a unified multi-camera service. The speed evaluation shows that lower analysis frequency can retain low vehicle-level error while narrowing measurement coverage. Rate selection should therefore consider required coverage alongside identity continuity and speed accuracy. Cameras monitoring fast passages or requiring more complete measurement may need higher rates, while others may tolerate reduced coverage in exchange for fewer inference inputs. This supports camera-specific allocation of a fixed inference budget rather than a uniform sampling policy.

Under the tested configuration, the ablations suggest that reliable sparse-frame tracking depends on both appearance evidence and the rules governing association decisions. The feature comparisons also separate the value of including a cue

from choosing its weight: a useful feature need not improve every operating point, and its inclusion does not establish an optimal mixing ratio. The design implication is not to discard motion, but to constrain its influence and combine guarded acceptance with repeated proposals and recovery opportunities, balancing identity consistency against missed associations.

Complementing the public benchmarks, qualitative inspection of confidential industrial CCTV recordings showed fewer visible identity switches than the alternative trackers examined during development. These recordings included turning, reversing, stationary vehicles, and partial occlusion. Integration into an industrial video-analytics service supports practical applicability; the recordings cannot be publicly released for confidentiality reasons.

7. CONCLUSION

AppearanceTracking leads the tested UA-DETRAC comparisons at 1–2 FPS. Controlled ablations identify complementary HSV and deep evidence, a sampling-dependent HOG benefit, and contributions from guarded association, iterative proposals, and bounded motion. Speed evaluation shows that lower sampling rates can retain low vehicle-level error when sufficient observations are available. However, increasingly sparse observations limit measurement coverage, making accuracy alone insufficient for assessing practical applicability. The framework combines appearance-guided association with measured geometry and timestamp-based speed, supporting sampling-rate selection according to both accuracy and required measurement coverage.

ACKNOWLEDGMENTS

This research was conducted as part of an industry-academia collaboration between POSTECH and LG Electronics.

REFERENCES

- Aharon, N., Orfaig, R., and Bobrovsky, B.-Z. (2022). BoT-SORT: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*.
- Bewley, A., Ge, Z., Ott, L., Ramos, F., and Upcroft, B. (2016). Simple online and realtime tracking. *Proceedings of the IEEE International Conference on Image Processing*, 3464–3468.
- Cao, J., Pang, J., Weng, X., Khirrodar, R., and Kitani, K. (2023). Observation-centric SORT: Rethinking SORT for robust multi-object tracking. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9686–9696.
- Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 886–893.
- Dubská, M., Sochor, J., and Herout, A. (2014). Automatic camera calibration for traffic understanding. *Proceedings of the British Machine Vision Conference*.
- Gao, R., Qi, J., and Wang, L. (2025). Multiple object tracking as ID prediction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27883–27893.
- Hartley, R. and Zisserman, A. (2004). *Multiple View Geometry in Computer Vision*, 2nd ed., Cambridge University Press.
- Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., and Leibe, B. (2021). HOTA: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 129, 548–578.
- Oquab, M., Darcet, T., Moutakanni, T., et al. (2024). DI-NOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193v2*.
- Sochor, J., Juránek, R., Špaňhel, J., Maršík, L., Šíroký, A., Herout, A., and Zemčík, P. (2018). Comprehensive data set for automatic single camera visual speed measurement. *IEEE Transactions on Intelligent Transportation Systems*.
- Wen, L., Du, D., Cai, Z., Lei, Z., Chang, M.-C., Qi, H., Lim, J., Yang, M.-H., and Lyu, S. (2020). UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. *Computer Vision and Image Understanding*, 193, 102907.
- Wojke, N., Bewley, A., and Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. *Proceedings of the IEEE International Conference on Image Processing*, 3645–3649.
- Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., and Wei, Y. (2022). MOTR: End-to-end multiple-object tracking with Transformer. *Proceedings of the European Conference on Computer Vision*.
- Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., and Wang, X. (2022). ByteTrack: Multi-object tracking by associating every detection box. *Proceedings of the European Conference on Computer Vision*, 1–21.
- Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., and Chen, J. (2024). DETRs beat YOLOs on real-time object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16965–16974.