Spectral-Aware Multi-Object Tracking in Harsh Aerial Perception Domains

Leandro Di Bella[1], Joseph Mimassi[1,2], Leon Denis[1], and Adrian Munteanu[1] *Member, IEEE*

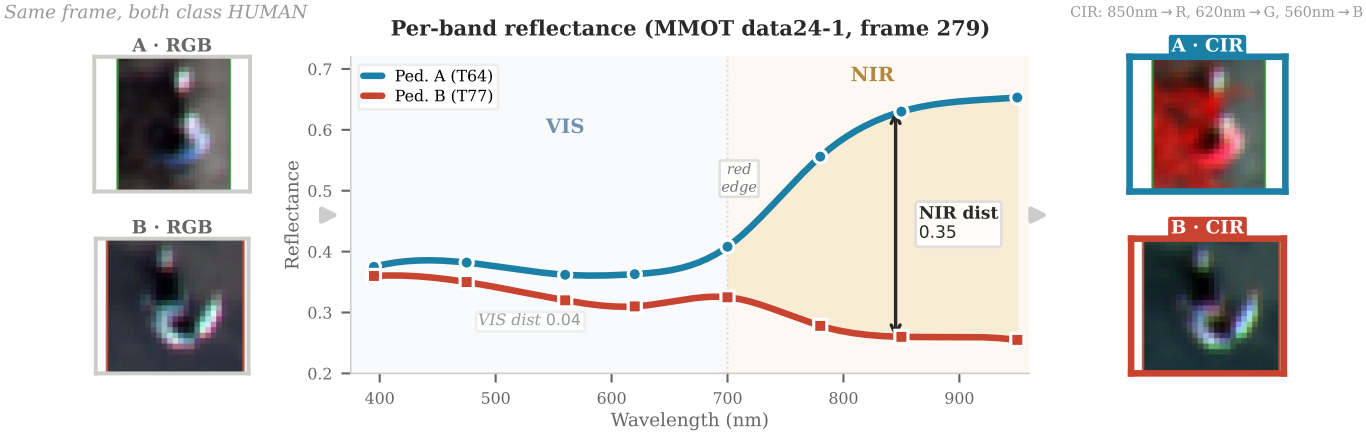

Fig. 1: **Spectral identity in action.** Two pedestrians from MMOT sequence, both wearing visually white clothing. The RGB views (left) are nearly indistinguishable (visible-spectrum (VIS) L2 distance 0.04). The per-band reflectance (centre) reveals why: A's curve climbs sharply at the red edge (∼700 nm), the canonical NIR signature of a synthetic dye absorbing only in the visible (e.g. polyester with a disperse azo black), while B's stays flat, characteristic of a broadband absorber (e.g. carbon-black cotton). The Color Infrared composites (CIR, right; 850 nm→R, 620 nm→G, 560 nm→B) make the 8× NIR divergence visible: A glows red, B remains dark. This material-chemistry difference is invisible to any RGB-only tracker but is exactly the signal SAT exploits at the association step.

***Abstract*—Perception in harsh robotics domains, including low-altitude aerial, underwater, and space, relies increasingly on multi-band sensors, yet the spectral signal is routinely discarded at the tracking stage. We argue this is a missed opportunity: material-dependent reflectance differences that are invisible in RGB imagery persist even when targets subtend only 10–20 pixels, making them the most reliable identity cue available in these regimes. We present SAT, a training-free tracker that injects an 8-band spectral descriptor directly into the data-association cost, with zero learned parameters in the spectral pathway. On the Multispectral Multi-Object Tracking (MMOT) benchmark (8 bands, 395–950 nm), SAT achieves Higher Order Tracking Accuracy (HOTA) 55.8 online (+2.2 over BoT-SORT) and 56.5 offline (+2.9), outperforming every training-time modification we tested. A formal analysis derives a $O(1/\sqrt{N})$ bound on the spectral sampling noise that predicts a sharp sensitivity cliff, and a descriptor sweep reveals that even an 8-D per-band mean captures the full spectral gain, suggesting a fundamental information limit when targets are small relative to sensor resolution.**

This work was supported by the Belgian Science Policy Office (BELSPO) and the Royal Higher Institute for Defence (RHID) within the research project ATTENTION. [1]Leandro Di Bella, Joseph Mimassi, Leon Denis and Adrian Munteanu are with the Department of Electronics and Informatics, Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels, Belgium. Leandro Di Bella, Leon Denis and Adrian Munteanu are also with IMEC, Kapeldreef 75, B-3001 Leuven, Belgium. [2]Joseph Mimassi is with the Robotics and Autonomous Systems Lab, Royal Military Academy of Belgium (RMA), Brussels, Belgium.

I. Introduction

Robotic perception in harsh field domains, including underwater (turbidity, wavelength-dependent attenuation, backscatter), space (extreme dynamic range, low photon counts, uncontrolled illumination), and high-altitude aerial (small targets, ego-motion, varying illumination), shares a common set of obstacles that break RGB-only pipelines. A recurring mitigation across all of these domains is **multi-band sensing**: multispectral and hyperspectral cameras on unmanned aerial vehicles (UAVs) and satellites, multibeam and side-scan sonar arrays underwater, thermal and event cameras on planetary rovers. Each modality provides additional channels of information that can disambiguate targets when standard imaging fails. Yet despite the availability of these rich sensor modalities, downstream perception components, particularly *tracking*, typically discard the multi-band signal, treating each frame as if it had been collapsed to grayscale. This paper argues that the association step of multi-object tracking is precisely where multi-band information should be injected, and demonstrates the principle on a challenging aerial benchmark representative of the small-target, high-ego-motion conditions shared across harsh-domain robotics.

We study this on the recently introduced MMOT benchmark [1], the first large-scale 8-band multispectral UAV multi-object tracking dataset (125 sequences, 488.8K oriented

annotations, 395–950 nm spanning visible and near-infrared). MMOT exhibits all the difficulties characteristic of harsh-domain perception: objects subtend only 10–20 pixels at typical UAV altitudes (80–200 m), the camera undergoes constant ego-motion from flight dynamics, illumination varies dramatically across sequences, and many target classes (pedestrians, cyclists) are spatially indistinguishable from each other in any single band.

Existing multispectral trackers either fuse the 8 bands early in a detection backbone (MOTRv2 [14] uses a 3D-Stem to merge bands before the ResNet backbone) or ignore them entirely at the association step: BoT-SORT [4], the current state of the art on UAV benchmarks, uses only 3 of 8 bands for camera-motion estimation and discards spectral information for identity matching. Prior multispectral tracking work focused on RGB-thermal pairs [2], [3] rather than full 8-band imagery, and standard appearance-based association via deep re-identification (ReID) features [6], [7] requires supervised training on RGB data. UCMCTrack [5] projects detections to a ground plane for camera-motion-invariant association but likewise ignores spectral information. Wasserstein distance has been used for multi-target assignment [15] but not for spectral histogram comparison in the association step. In short, no existing MOT method exploits the multi-band signal at the point where it matters most for identity: the data association.

This is a fundamental missed opportunity. An object's spectral signature, i.e. how it reflects light across 8 wavelengths, is an identity-discriminative feature that is *orthogonal* to spatial appearance and *invariant* to viewpoint, and therefore particularly valuable when spatial features are degenerate, as they typically are in harsh-domain imagery. Two pedestrians wearing identical-looking clothing appear indistinguishable in visible bands but exhibit distinct near-infrared (NIR) reflectance from different fabric materials: cotton absorbs strongly at 850 nm while polyester reflects (Fig. 1). The same material-dependent principle applies to vehicles (paint chemistry), bicycles (metal-rubber-plastic composition), and, crucially for the broader harsh-domain community, to underwater targets distinguished by multi-frequency sonar return characteristics (metallic vs. organic vs. sediment), space targets distinguished by thermal emissivity signatures (solar panels vs. insulation blankets vs. bare metal), and subterranean objects distinguished by LiDAR intensity profiles. The common thread is that *spatial appearance degrades under harsh conditions while spectral identity persists*, and this paper demonstrates a principled, lightweight way to exploit that complementarity at the association stage.

We present **Spectral-Aware Tracking (SAT)**, a training-free framework that injects multi-band information directly into the association step. Our contributions:

1) **Spectral identity belongs at association, not training.** We show that injecting a simple training-free spectral descriptor, the Multispectral Appearance Signature (MAS) compared via the Wasserstein-1 distance, at the association step outperforms every training-time architectural modification we tested. Zero learned parameters in the spectral pathway.
2) **A fundamental limit on spectral-spatial fusion in small-target regimes.** We derive that the per-frame spectral sampling noise scales as $\sigma_W(N) = O(1/\sqrt{N})$, bounding the optimal fusion weight from above as a function of object pixel count.
3) **State-of-the-art results on MMOT with zero spectral parameters.** SAT achieves HOTA 55.8 online (+2.2 over BoT-SORT, Association Accuracy (AssA) 64.6 to 70.6), with gains concentrated on small, visually ambiguous classes.

II. Method

SAT (Fig. 2) builds on BoT-SORT [4] by injecting a spectral identity term into the association cost. The core observation is that existing trackers discard multi-band information precisely where it would be most useful: detection networks learn to fuse 8 bands for *what* is present, but the tracking step matches objects using only spatial overlap (IoU), as if all bands had been collapsed into grayscale. We argue that the spectral reflectance profile of an object, its unique pattern of absorption and reflection across 8 wavelengths from 395 to 950 nm, constitutes an identity-discriminative fingerprint that is complementary to spatial features and invariant to viewpoint changes common in UAV flight.

A natural alternative would be to inject spectral information at training time, modifying the detection backbone or the end-to-end tracker to be spectrally aware. The MMOT baseline [1] already does this at the detection stage: YOLO11 [18] is retrained on all 8 channels, and MOTRv2 adds a Spectral 3D-Stem (ConvMSI) to fuse bands before the ResNet backbone. These adaptations are effective for detection, but the spectral signal is still discarded at association. This motivates our inference-time design: rather than modifying the trained pipeline, we inject spectral identity *after* it, in the association step, where it cannot corrupt learned representations.

The full SAT pipeline is illustrated in Fig. 2. An 8-band multispectral image ($H \times W \times 8$, wavelengths 395–950 nm) is first processed by a YOLO11L [18] detector trained on all 8 channels, producing oriented bounding boxes. For each detection, we extract a per-band intensity histogram from the bounding-box crop and L2-normalize the concatenation into a 64-dimensional Multispectral Appearance Signature (MAS). In parallel, a Kalman filter predicts each existing track's state forward, and sparse-optical-flow camera-motion compensation corrects for UAV ego-motion. The association step then builds a fused cost matrix that combines the standard rotated intersection-over-union (IoU) spatial cost with the Wasserstein Spectral Distance (WSD) between each track's temporally-smoothed MAS and each detection's MAS. Hungarian matching [20] on this fused cost produces the optimal track-to-detection assignment, after which track states are updated, new tracks are initialised, and lost tracks are removed. The entire online pipeline runs causally in a single forward pass at $\sim$45 frames per second (FPS) on CPU.

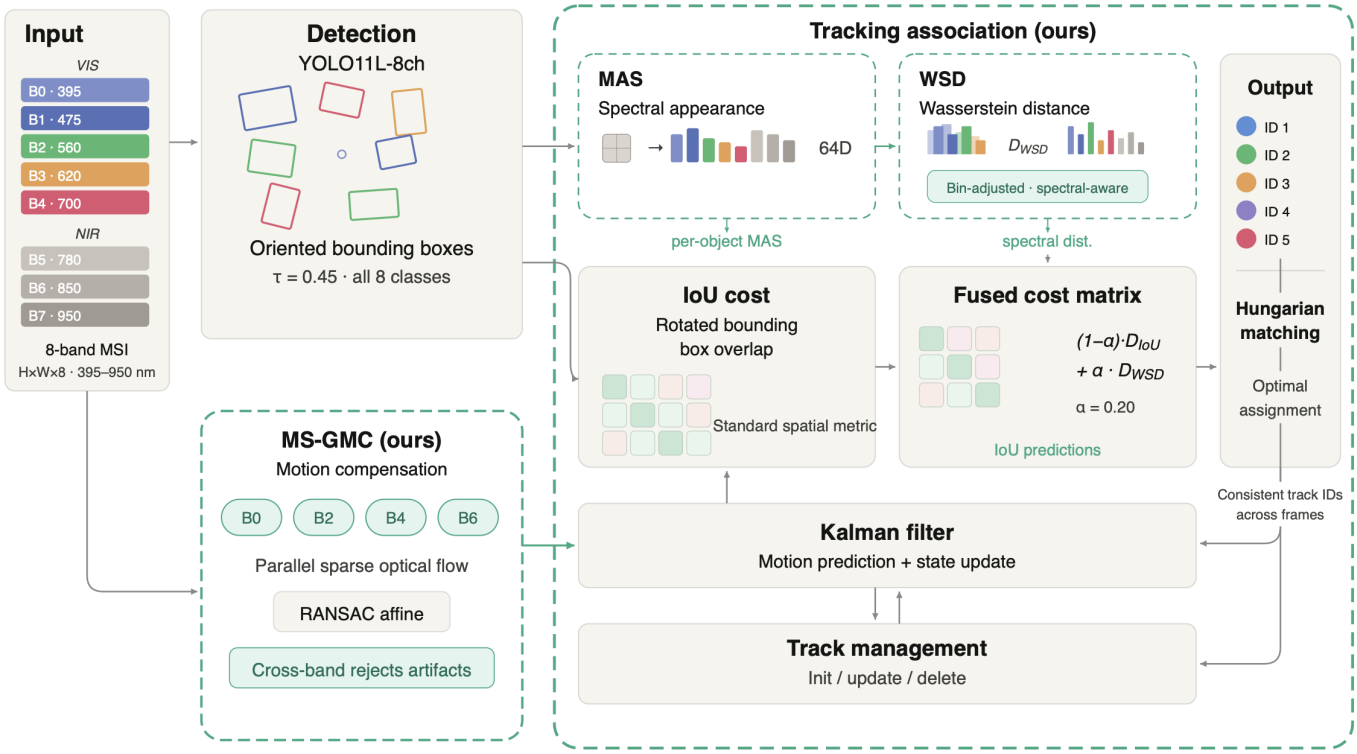

Fig. 2: **SAT pipeline.** An 8-band multispectral image (395–950 nm) is processed by an 8-channel YOLO11L [18] detector. For each detection, we extract per-band intensity histograms and L2-normalize their concatenation into a 64-D Multispectral Appearance Signature (MAS). The Wasserstein Spectral Distance (WSD) compares track and detection signatures via per-band cumulative distribution function (CDF) differences, and is fused with rotated intersection-over-union (IoU) into a single association cost. The result is a training-free, drop-in upgrade to the BoT-SORT pipeline that injects spectral identity directly into the matching step.

A. Spectral Descriptor and Distance

For each detection d_j with oriented bounding box, we extract the axis-aligned crop from the raw C-band image $\mathbf{I} \in \mathbb{R}^{H\times W\times C}$. Let $\mathcal{R} \subset \{1,\dots,H\}\times\{1,\dots,W\}$ denote the set of $N = |\mathcal{R}|$ pixel locations inside the crop. For each spectral band $b \in \{0,\dots,C{-}1\}$, we partition the intensity range $[0, 255]$ into B uniform bins and count the pixels falling in each:

$$\hat{h}_b(k) = \left|\left\{(x,y) \in \mathcal{R} \;:\; I_b(x,y) \in \mathrm{bin}_k\right\}\right|, \quad k \in \{1,\dots,B\} \tag{1}$$

The normalised per-band histogram $h_b \in \mathbb{R}^B$ divides by the total pixel count:

$$h_b = \frac{\hat{h}_b}{\sum_{k=1}^{B} \hat{h}_b(k)} = \frac{\hat{h}_b}{N} \tag{2}$$

Each h_b is non-negative and sums to one, representing the empirical intensity distribution for band b. The **Multispectral Appearance Signature (MAS)** is their concatenation:

$$\boldsymbol{s} = [h_0;\ h_1;\ \dots;\ h_{C-1}] \ \in \ \mathbb{R}^{CB} \tag{3}$$

Each B-element slice of $\boldsymbol{s}$ is a valid probability distribution by construction. Track-level signatures are maintained via an exponential moving average (EMA):

$$\bar{\boldsymbol{s}}_t = \beta\, \bar{\boldsymbol{s}}_{t-1} + (1{-}\beta)\, \boldsymbol{s}_t, \quad \beta = 0.9 \tag{4}$$

Because the EMA is a convex combination, each per-band slice of $\bar{\boldsymbol{s}}_t$ remains a valid probability distribution, so no re-normalisation is needed before computing the Wasserstein distance. The retention weight $\beta = 0.9$ means each new observation contributes only 10% to the running signature: a single noisy frame (glare, partial occlusion) corrupts at most 10% of the descriptor, at the cost of requiring $\sim 1/(1{-}\beta) = 10$ frames for a new track's signature to stabilise. On MMOT, $C{=}8$ bands and $B{=}8$ bins, yielding a 64-D descriptor.

a) Wasserstein Spectral Distance (WSD).: To compare MAS descriptors we use the **Wasserstein-1 distance** (Earth Mover's Distance, EMD). Cosine distance treats each histogram bin as an independent dimension, so a peak that migrates from bin k to bin $k{+}1$ (minor illumination change) produces the same distance as a shift to bin $k{+}(B{-}1)$ (genuinely different material). Wasserstein respects the ordinal structure of intensity bins, producing a distance proportional to the shift *magnitude*. For two probability vectors $\mathbf{p}, \mathbf{q} \in \Delta^B$ over B ordered bins with $\Delta^B = p \in \mathbb{R}^B : p_i \geq 0, ; \sum_i p_i = 1$, the Wasserstein-1

distance admits a closed-form solution via the cumulative distribution function (CDF) difference [15]:

$$W_1(\mathbf{p}, \mathbf{q}) = \sum_{k=1}^{B} \big| F_p(k) - F_q(k) \big| \tag{5}$$

where $F(k) = \sum_{l=1}^{k} p(l)$ is the CDF.

Since each B-element slice of $\bar{\boldsymbol{s}}$ is a probability vector, the WSD between a track signature $\bar{\boldsymbol{s}}_i$ and a detection signature $\boldsymbol{s}_j$ is simply the per-band Wasserstein average:

$$D_{\text{WSD}}(\bar{\boldsymbol{s}}_i, \boldsymbol{s}_j) = \frac{1}{C(B-1)} \sum_{b=0}^{C-1} W_1\big(\bar{\boldsymbol{s}}_i^{(b)},\ \boldsymbol{s}_j^{(b)}\big) \tag{6}$$

where $\bar{\boldsymbol{s}}_i^{(b)} = \bar{\boldsymbol{s}}_i\big[bB{:}(b{+}1)B\big]$ denotes the b-th band slice and the division by $C(B-1)$ normalises to $[0, 1]$.

The fused association cost for track T_i and detection d_j combines the spatial and spectral terms:

$$D_{ij} = (1 - \alpha)\, D_{\text{IoU}}(T_i, d_j) + \alpha\, D_{\text{WSD}}(\bar{\boldsymbol{s}}_i, \boldsymbol{s}_j) \tag{7}$$

applied in the first-stage (high-confidence) association of the BoT-SORT cascade with $\alpha = 0.20$.

B. Multispectral Global Motion Compensation (MS-GMC)

UAV platforms introduce persistent ego-motion from flight dynamics (pan, tilt, altitude changes) that corrupts motion-based association. Standard camera motion compensation estimates frame-to-frame affine transforms from sparse optical flow on a single grayscale channel. This is fragile in conditions where the chosen band has low scene contrast: overexposed VIS bands in direct sunlight, dark VIS bands in shadow regions, or NIR-dark water surfaces.

We propose **Multispectral GMC (MS-GMC)**: parallel sparse optical flow computation across K spectral bands, with cross-band correspondence fusion for robust affine estimation. For each selected band b_k, we independently detect keypoints and track them with pyramidal Lucas-Kanade optical flow. All valid correspondences are concatenated into a single pool:

$$\mathcal{P} = \bigcup_{k=1}^{K} \{(\mathbf{p}_{\text{prev}}^{b_k}, \mathbf{p}_{\text{curr}}^{b_k})\} \tag{8}$$

from which a single 4-DOF affine transform (rotation, translation, scale) is estimated via RANSAC. By drawing correspondences from both VIS (bands 0, 2) and NIR (bands 4, 6), MS-GMC maintains keypoint density even when individual spectral bands are featureless, NIR provides texture in shadow regions where VIS fails, and VIS provides structure in NIR-dark water/glass regions. The cross-band fusion inherently rejects band-specific artifacts (sensor noise patterns, spectral blooming) as RANSAC outliers, since these produce inconsistent correspondences across bands.

C. Tracker Configuration and Offline Refinement

Beyond the spectral descriptor, we implemented a class-consistency penalty discourages cross-superclass associations, and rotated-IoU non-maximum suppression (NMS) removes near-duplicate oriented boxes. We also propose an offline alternative with two optional refinements operating on saved track files without re-running the detector or the spectral descriptor: (i) linear interpolation filling ≤ 2-frame gaps [6], [10], and (ii) Bidirectional Spectral Track Refinement (BSTR). BSTR re-runs SAT on the time-reversed video, splits each forward tracklet at spectral discontinuities (consecutive-frame WSD above a threshold), and re-attaches the resulting fragments to backward tracklets via Hungarian matching on a cost combining spatial proximity, class consistency, and WSD between fragment-mean signatures. The forward-only and backward-only views thus arbitrate each other's identity decisions, fixing ID switches a causal tracker cannot undo.

III. Experiments

We evaluate on the MMOT benchmark [1] (8-band, 125 sequences, 488.8K oriented bounding box (OBB) annotations, 3 superclasses; 50-sequence test split). All methods use the same YOLO11L [18] detector trained on 8-band MMOT data and provided by [1]; results for prior trackers (SORT through BoT-SORT) are taken directly from the benchmark paper under this shared detector. Our proposed method reuse exactly the same detector checkpoint and detections as our baselines. Our primary metric is HOTA [8].

A. Results

Table I shows the main results. SAT online achieves **HOTA 55.8** ($+2.2$ over BoT-SORT), with the gain dominated by association quality (AssA $64.6 \rightarrow 70.6$ compared to BoT-SORT); Detection Accuracy (DetA) is unchanged (45.7), confirming that spectral appearance features aid identity discrimination rather than detection. The learned 128-D ReID encoder, a small convolutional neural network (CNN) ($\sim$210k params) trained with triplet loss on 8-channel MMOT crops, achieves the same HOTA as the analytical histogram at substantially higher inference cost, showing that a training-free descriptor already captures the available spectral signal. Stacking BSTR and interpolation yields **HOTA 56.5** ($+2.9$, $+8.1$ AssA, $+4.8$ IDF1); the two offline steps are complementary because they target different failure modes: BSTR fixes ID switches via bidirectional spectral consensus while interpolation recovers transient detection gaps.

The spectral gain is concentrated on classes most affected by spatial appearance ambiguity: HUMAN gains $+5.8$ AssA, BIKE $+12.6$, precisely the targets that appear as small, visually-similar blobs at UAV altitude but exhibit distinct NIR reflectance from different clothing materials and surface compositions (Fig. 1). VEHICLE gains only $+1.3$ because vehicles are large enough for spatial discrimination alone. This per-class pattern directly validates the spectral fingerprint hypothesis: NIR-dependent classes gain the most, because their identity signal is spectral rather than spatial. The spectral

TABLE I: Main results on the MMOT benchmark. All methods use the same YOLO11L-8ch [18] detector. Class-averaged metrics weight each of the 3 superclasses equally; detection-averaged metrics weight each detection equally. Best online result in **bold**, second best are in underlined, best overall *underlined*. ‡ uses a learned ReID encoder instead of our training-free histogram descriptor. MOTA: Multiple Object Tracking Accuracy; IDF1: Identity F1 score.

	Class-averaged					**Detection-averaged**				
Method	HOTA	DetA	AssA	MOTA	IDF1	HOTA	DetA	AssA	MOTA	IDF1
Online										
SORT [9]	27.2	25.7	30.0	24.3	29.1	35.0	27.6	44.8	25.7	33.7
ByteTrack [10]	40.5	37.0	46.2	34.2	44.1	46.0	41.9	51.5	37.8	46.7
OC-SORT [11]	29.5	27.3	32.8	25.1	31.9	37.5	29.5	48.0	27.5	37.0
MOTR [12]	39.0	27.1	60.1	26.5	44.6	48.4	35.4	68.4	32.2	54.7
MeMOTR [13]	42.3	29.3	66.3	31.3	45.9	50.9	37.1	70.9	40.8	56.0
MOTRv2 [14]	49.2	37.8	67.7	43.1	57.3	54.5	44.1	68.8	50.9	64.6
BoT-SORT [4]	53.6	**45.7**	64.6	46.2	61.0	60.7	55.0	68.7	59.4	69.4
SAT, learned ReID‡ (ours)	55.8	**45.7**	70.4	49.4	64.7	63.1	56.3	72.6	63.8	74.0
SAT (ours)	**55.9**	**45.7**	**70.6**	**49.5**	**64.9**	**63.2**	**56.3**	**72.7**	**63.9**	**74.2**
Offline post-processing on SAT output										
SAT + interp	56.1	45.7	71.6	49.2	64.7	63.1	56.1	72.6	63.4	73.7
SAT + BSTR	56.2	*45.8*	71.4	*49.7*	*65.9*	*63.7*	*56.0*	74.1	*64.1*	*75.8*
SAT + BSTR + interp	*56.5*	45.6	*72.7*	49.2	65.8	63.5	55.6	*74.2*	63.2	75.5

TABLE II: Per-class breakdown. Spectral features disproportionately benefit HUMAN and BIKE, the classes most affected by spatial appearance ambiguity at UAV altitude.

	HOTA			AssA		
Class	Base	SAT	Δ	Base	SAT	Δ
HUMAN	42.4	**45.4**	**+3.0**	40.5	**46.3**	**+5.8**
VEHICLE	78.8	79.0	+0.2	84.1	85.4	+1.3
BIKE	42.1	**44.4**	**+2.2**	51.5	**64.1**	**+12.6**

TABLE III: Descriptor information floor. The 8-D per-band mean recovers the full spectral gain; richer descriptors add no measurable signal. ♭NDI: illumination-invariant normalised differences.

Spectral feature	Dim.	HOTA	Δ
None (IoU only)	0	53.63	—
Histogram, B=8	64	55.86	+2.23
Histogram, B=32	256	55.87	+2.24
Learned 128-D ReID	128	55.77	+2.13

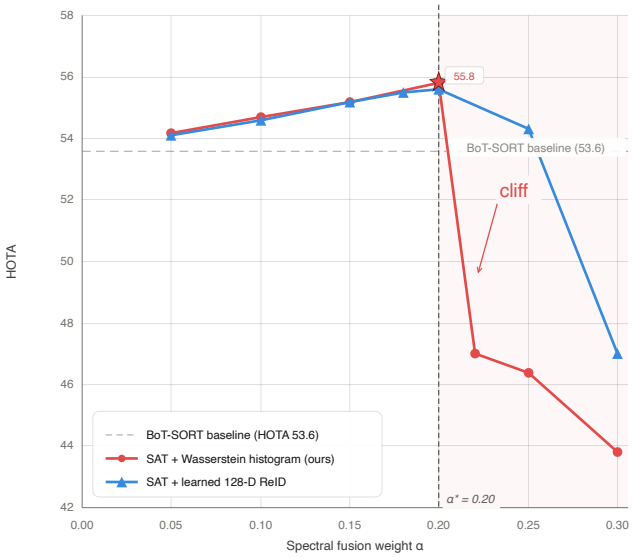

Fig. 3: **Sensitivity cliff.** All descriptors peak at α=0.20 and degrade beyond.

histogram captures material-dependent reflectance differences invisible to RGB: cotton absorbs strongly at 850 nm while polyester reflects, enabling disambiguation of identically-dressed pedestrians; automotive paints produce characteristic VIS-NIR responses; and metal, rubber, and plastic bicycle components have distinct NIR absorption profiles. These signatures are inherently identity-discriminative and complementary to geometric (IoU) matching.

B. Analysis

A natural question is how much spectral structure is actually needed. Table III shows a perhaps-surprising answer: the 8-D per-band mean already recovers +2.21 HOTA, virtually the entire spectral gain. The learned encoder (contrastive-trained with triplet loss on 64×64 8-channel crops from MMOT-train with hard same-class negatives) does not break the ceiling despite having ∼210k learnable parameters, confirming that the bottleneck is the data's information content rather than the descriptor's capacity.

At MMOT crop sizes, only the lowest-order moments of the per-band distribution are reliably estimable; richer descriptors

provide capacity that cannot be filled with reliable signal.

This information floor is explained by a sharp *sensitivity cliff* at $\alpha > 0.20$ (Fig. 3): HOTA drops from 54.6 to 38.4 for histograms, a 16.2-point collapse. The cliff is universal across every descriptor tested, though its severity varies (learned encoders degrade gracefully at -1.6). The root cause is quantifiable. An object covering N pixels yields empirical histograms $\hat{p}_b \sim \frac{1}{N}\text{Multinomial}(N, p_b)$, and the expected Wasserstein noise between two samples of the *same* object is bounded by $\sigma_W(N) \leq 1/\sqrt{C\pi N}$ [19]. The fused cost ranks a true match below a false match when $(1-\alpha)(D_{\text{IoU}}^{\text{false}} - D_{\text{IoU}}^{\text{true}}) + \alpha(\sigma_W(N) - \Delta_{\text{spec}}) < 0$, defining a critical pixel count $N_{\text{crit}} = 1/(C\pi\Delta_{\text{spec}}^2)$ below which spectral noise exceeds signal. The formula has one free parameter, Δ_{spec}, the typical spectral distance between the closest same-class false match. Fitting it to the observed cliff position on MMOT (C=8) gives $\Delta_{\text{spec}} \approx 0.045$, yielding $N_{\text{crit}} \approx 20\,\text{px}$; with a one-$\sigma_W$ safety margin, $N_{\text{crit}} \sim 80\,\text{px}$, which coincides with the 10th percentile of the observed pixel-count distribution. The cliff is thus governed by the small-N tail, not its median.

We ruled out three alternative explanations: **(F1)** an illumination-invariant NDI descriptor [17] (28-D pairwise normalised band differences) hits the same cliff, so it is not an illumination effect; **(F2)** a spectral memory bank for long-term re-ID produced more false re-IDs than true recoveries at every threshold (-0.01 to -1.3 HOTA), since the gallery decision requires tighter discrimination than frame-to-frame matching; **(F3)** Vertex Component Analysis (VCA) [21] + Fully Constrained Least Squares (FCLS) [22] spectral unmixing did not exceed the per-band mean, since mixture estimation is itself high-variance at small N. All three are consistent with the sampling-noise bound.

On MMOT, this analysis implies that improving spectral descriptors beyond the current state of the art will require either (i) higher-resolution imagery (lower altitude or better sensors), (ii) temporal aggregation of spectral features across multiple frames so that the effective N grows linearly with the tracklet length, or (iii) descriptor designs that down-weight pairs with $N \leq N_{\text{crit}}$ in the cost matrix. Of these, (ii) is the most actionable: our EMA over per-frame histograms is the simplest temporal aggregation, and it is what allows mature tracks to tolerate per-frame noise that would kill new tracks with a single observation. We suggest that the same $1/\sqrt{N}$ scaling could govern spectral-spatial fusion in any harsh-domain multi-band tracking problem, whether the "bands" are optical wavelengths, sonar frequencies, or thermal channels, whenever targets are small relative to the sensor resolution. Whether this holds is a one-line empirical test: sweep α and fit N_{crit} against the pixel-count distribution on the new dataset.

IV. Conclusion and Future Work

We presented SAT, a training-free framework that injects multi-band signal into the association step of a multi-object tracker. Three findings emerge. First, a *design principle*: on the MMOT benchmark, spectral identity is more effective at association than at training which achieves HOTA 55.8 online (+2.2+2.2 +2.2) and 56.5 offline (+2.9+2.9 +2.9) with zero learned parameters in the spectral pathway. Second, a *fundamental limit*: the per-frame spectral sampling noise scales as $O(1/\sqrt{N})$, bounding the optimal fusion weight from above and producing a sharp sensitivity cliff that appears governed by the small-N tail of the pixel-count distribution rather than by illumination or descriptor sophistication. Third, a *notable empirical observation*: the 64-D histogram matches every richer descriptor we tested at the optimal operating point, suggesting that only the lowest-order spectral moments carry reliably estimable signal at typical UAV crop sizes.

We see several directions for future work, motivated by the cross-domain perspective of this workshop. First, the $1/\sqrt{N}$ scaling law that governs the sensitivity cliff should be validated on other harsh-domain multi-band datasets; RGB-thermal tracking benchmarks (RGBT234, LasHeR) are a natural next step, and marine sonar tracking datasets would test whether the same principle extends beyond electromagnetic spectra to acoustic modalities. Second, temporal spectral aggregation beyond our simple EMA deserves investigation: a Bayesian update scheme that explicitly models per-frame illumination noise could lower the effective σ_W and push the optimal α higher, especially for long-lived tracks. Third, the sensitivity cliff suggests an adaptive fusion strategy where α is modulated per-pair as a function of the detection's pixel count N: large objects could safely absorb a higher spectral weight while small objects default to IoU-dominated association. Fourth, the success of the training-free approach raises the question of whether spectral association can be combined with learned features in a regime-aware manner: using learned descriptors for large targets (where they have enough pixels to be reliable) and analytical descriptors for small targets (where simplicity wins). Finally, we hope the framing offered here, viewing UAV multispectral tracking as an instance of a broader harsh-domain multi-band perception problem, encourages cross-research with the marine and space tracking communities, where similar small-target, degraded-appearance challenges arise from sonar, thermal, and hyperspectral sensors.

References

[1] Authors of MMOT, "MMOT: A Large-Scale Benchmark for Multi-Object Tracking in Multispectral UAV Imagery," *NeurIPS Datasets and Benchmarks*, 2025.

[2] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "RGB-T object tracking: Benchmark and baseline," *Pattern Recognition*, vol. 96, p. 106977, 2019.

[3] H. Zhang *et al.*, "Multi-modal multi-object tracking," *CVPR*, 2022.

[4] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "BoT-SORT: Robust associations multi-pedestrian tracking," *arXiv:2206.14651*, 2022.

[5] K. Yi *et al.*, "UCMCTrack: Multi-object tracking with uniform camera motion compensation," *AAAI*, 2024.

[6] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," *ICIP*, pp. 3645–3649, 2017.

[7] Y. Zhang *et al.*, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," *IJCV*, vol. 129, no. 11, pp. 3069–3087, 2021.

[8] J. Luiten *et al.*, "HOTA: A higher order metric for evaluating multi-object tracking," *IJCV*, vol. 129, no. 2, pp. 548–578, 2021.

[9] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," *ICIP*, pp. 3464–3468, 2016.

[10] Y. Zhang *et al.*, "ByteTrack: Multi-object tracking by associating every detection box," *ECCV*, 2022.

[11] J. Cao *et al.*, "Observation-centric SORT: Rethinking SORT for robust multi-object tracking," *CVPR*, 2023.
[12] F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei, "MOTR: End-to-end multiple-object tracking with transformer," *ECCV*, 2022.
[13] R. Gao and L. Wang, "MeMOTR: Long-term memory-augmented transformer for multi-object tracking," *CVPR*, 2024.
[14] Y. Zhang *et al.*, "MOTRv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors," *CVPR*, 2023.
[15] G. Peyré and M. Cuturi, "Computational optimal transport," *Foundations and Trends in Machine Learning*, vol. 11, no. 5–6, pp. 355–607, 2019.
[16] F. A. Kruse *et al.*, "The Spectral Image Processing System (SIPS)," *Remote Sensing of Environment*, vol. 44, no. 2–3, pp. 145–163, 1993.
[17] J. W. Rouse, R. H. Haas, J. A. Schell, and D. W. Deering, "Monitoring vegetation systems in the Great Plains with ERTS," in *Proc. 3rd ERTS Symposium*, NASA SP-351, vol. 1, pp. 309–317, 1974.
[18] G. Jocher and J. Qiu, "Ultralytics YOLO11," 2024. https://github.com/ultralytics/ultralytics.
[19] N. Fournier and A. Guillin, "On the rate of convergence in Wasserstein distance of the empirical measure," *Probability Theory and Related Fields*, vol. 162, no. 3–4, pp. 707–738, 2015.
[20] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1–2, pp. 83–97, 1955.
[21] J. M. P. Nascimento and J. M. B. Bioucas-Dias, "Vertex component analysis: A fast algorithm to unmix hyperspectral data," *IEEE Trans. Geoscience and Remote Sensing*, vol. 43, no. 4, pp. 898–910, 2005.
[22] D. C. Heinz and C.-I Chang, "Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery," *IEEE Trans. Geoscience and Remote Sensing*, vol. 39, no. 3, pp. 529–545, 2001.