# Recovering Origin–Destination Flows from Bus CCTV: Early Results from Nairobi and Kigali

**Nthenya Kyatha**
University of Massachusetts Amherst
Amherst, MA, USA
mkyatha@umass.edu

**Jay Taneja**
University of Massachusetts Amherst
Amherst, MA, USA
jtaneja@umass.edu

## Abstract

Public transport in sub-Saharan Africa (SSA) often operates in overcrowded conditions where existing automated systems fail to capture reliable passenger flow data. Leveraging onboard CCTV already deployed for security, we present a baseline pipeline that combines YOLOv12 detection, BotSORT tracking, OSNet embeddings, OCR-based timestamping, and telematics-based stop classification to recover bus origin–destination (OD) flows. On annotated CCTV segments from Nairobi and Kigali buses, the system attains high counting accuracy under low-density, well-lit conditions (recall $\approx$95%, precision $\approx$91%, F1 $\approx$93%). It produces OD matrices that closely match manual tallies. Under realistic stressors such as overcrowding, color-to-monochrome shifts, posture variation, and non-standard door use, performance degrades sharply (e.g., $\sim$40% undercount in peak-hour boarding and a $\sim$17 percentage-point drop in recall for monochrome segments), revealing deployment-specific failure modes and motivating more robust, deployment-focused Re-ID methods for SSA transit.

## 1 Introduction

Public transport in sub-Saharan Africa (SSA) is dominated by buses and matatus that frequently operate well beyond capacity. Reliable Origin–Destination (OD) matrices are critical for route design, scheduling, equity analysis, and assessing the impacts of electric fleet adoption, yet most agencies still rely on costly manual surveys that provide only partial and infrequent insight Zalewski et al. [2019].

A range of automated technologies has been explored for OD estimation, including Automated Fare Collection (AFC) systems, Automated Passenger Counters (APCs), RFID-based boarding systems, Bluetooth, and mobile phone data, as well as multi-source approaches combining GIS data with IC card records Jafari Kang et al. [2021], Liu et al. [2021], González et al. [2020], Ozbay et al. [2017], Larijani et al. [2015], Kong et al. [2021]. While effective in high-resource contexts, these methods suffer from high cost, maintenance overhead, limited coverage, and privacy concerns, and are rarely deployed at scale in SSA bus networks. They also typically provide only aggregate boarding and alighting counts, making it difficult to track individuals through the vehicle or distinguish official stops from opportunistic roadside pickups.

Video-based re-identification (Re-ID) offers a complementary path by leveraging existing onboard CCTV. Prior work has combined YOLO detection with Re-ID features for passenger matching Shimada et al. [2019], or focused on head-only features and pose-guided attention Zhao et al. [2022], Miao et al. [2019], but these approaches are brittle under occlusion, posture changes, and clothing similarity. TransitReID Huang et al. [2025] introduces occlusion-resistant embeddings and hierarchical dynamic matching tailored to transit settings, yet it assumes relatively controlled, higher-resolution, predominantly color footage and does not account for illegal roadside stops or

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: UrbanAI: Harnessing Artificial Intelligence for Smart Cities.

frequent color-to-monochrome switching. As a result, it has not been evaluated in the low-resolution, modality-shifting, overcrowded CCTV environments typical of SSA buses.

These gaps motivate our baseline pipeline for OD inference in SSA buses. Our design leverages existing CCTV and telematics streams without additional hardware, combining passenger detection and tracking, OCR-based timestamping, Region-of-Interest (ROI) based event counting at doors, and cross-camera association with stop classification. This four-stage architecture (detailed in Sec. 2) produces stop-level OD matrices while surfacing deployment-specific challenges. In contrast to prior work, we deploy the system directly on Nairobi and Kigali buses, demonstrating feasibility under real-world conditions and highlighting challenges unique to SSA transit, such as extreme overcrowding, non-standard door usage, frequent color-to-monochrome CCTV transitions, and illegal roadside stops that remain underexplored in existing literature.

## 2 System Design

Our baseline system integrates four stages (Fig. 2 in Appendix A.2): (i) per-camera passenger detection and local tracking using YOLOv12 with BotSORT and Re-ID features; (ii) OCR-based timestamp extraction, combined with the known frame rate (FPS) to align events at one-second resolution; (iii) Region-of-Interest (ROI) based IN/OUT counting at each door; and (iv) cross-camera association and telematics alignment, which link front- and exit-camera tracklets where necessary, classify stop locations, and ultimately construct OD matrices.

### 2.1 Detection and local tracking

For both Cam-A and Cam-B, passengers are detected in every frame using YOLOv12, chosen for its balance of accuracy and speed on crowded, low-resolution footage. BotSORT maintains temporal consistency by assigning local IDs to each detection, augmented with appearance features from a lightweight Re-ID backbone (OSNet). This produces per-camera tracklets that capture each passenger's short-term motion through the scene. Importantly, both cameras are treated as bidirectional sensors: each ROI crossing may correspond to either a boarding or an alighting, since real-world behavior in Nairobi and Kigali often diverges from the intended "front-in, rear-out" design. These per-camera tracklets form the basis for both local counts and cross-camera associations.

### 2.2 Cross-camera Re-Identification

Tracklets from Cam-A (front aisle) and Cam-B (exit door) are then associated using appearance embeddings. Let $\mathbf{e}_i$ and $\mathbf{e}_j$ denote the OSNet embeddings for tracklets $i$ and $j$ from Cam-A and Cam-B, respectively. We compute cosine distance

$$d_{ij} = 1 - \frac{\mathbf{e}_i^\top \mathbf{e}_j}{\|\mathbf{e}_i\|_2 \, \|\mathbf{e}_j\|_2}. \tag{1}$$

These distances populate a cost matrix $D = [d_{ij}]$ over active tracklets in the two views. We apply the Hungarian algorithm to obtain a minimum-cost bipartite matching, subject to a gating threshold $d_{ij} \leq \tau_{\text{reid}}$; pairs that exceed the threshold remain unmatched and start new global trajectories. This cross-camera Re-ID step links a passenger's trajectory across the two streams, enabling reconstruction of complete journeys from boarding at the entrance to alighting at the exit and transitioning from per-camera IDs to global passenger trajectories needed for OD matrix construction.

### 2.3 OCR Timestamp Extraction and Aggregation

Each CCTV frame contains a time overlay produced by the recorder. To synchronize vision-based events with telematics data, we apply OCR to extract these timestamps. Consecutive frames often share the same overlay value, so we combine the OCR strings with the known frame rate (FPS) to generate a reliable per-second timeline. Boarding and alighting events detected from ROI crossings are aggregated into per-second IN/OUT counts, producing a time-stamped event log that aligns with bus stop arrivals.

### 2.4 Baseline ROI counting

In the baseline configuration, ROI-based counting at each door treats every tracked box that crosses the ROI as a candidate boarding or alighting event, independent of the door state. This simple strategy works well when passenger density is low and few people linger near the doors, but it produces false positives when riders stand or move near the exit while the bus is in motion, or when conductors repeatedly enter and leave the ROI without actually alighting. These errors are most pronounced at the exit door (Cam-B), where passengers frequently cluster around the steps.

### 2.5 Telematics Integration

The telematics stream provides vehicle data such as GPS coordinates, wheel-based speed, odometer reading, voltage, current, and various vehicle states. In this work, we primarily use wheel-based speed, odometer reading, and GPS to estimate distance along the route and identify potential stop periods. In the simplest baseline formulation, candidate stops are detected whenever the bus comes to a complete halt (wheel speed $v_t = 0$) and the GPS position lies near a known stop coordinate.

Let $v_t$ denote wheel-based speed at time $t$, $g_t$ the GPS coordinate, and $\mathcal{S}$ the set of official stop coordinates. We define the distance to the nearest stop as

$$\mathrm{dist}(g_t, \mathcal{S}) = \min_{s \in \mathcal{S}} \|g_t - s\|_2. \tag{2}$$

We label an *official stop* if $v_t = 0$ and $\mathrm{dist}(g_t, \mathcal{S}) \leq \delta_{\text{gps}}$. To better capture real-world behavior, we extend this baseline with a more flexible notion of *illegal stops*: we label an illegal stop if either (i) $v_t = 0$ and $\mathrm{dist}(g_t, \mathcal{S}) > \delta_{\text{gps}}$, or (ii) $0 < v_t < \tau_{\text{slow}}$ and the door-status signal indicates an open door, coinciding with ROI events. These rules capture passengers boarding or alighting amidst congestion and at unscheduled roadside points. By combining ROI-based events with this stop classification, the system distinguishes official passenger exchanges at bus stops from opportunistic exchanges during slow traffic or at off-stop locations.

### 2.6 OD matrix construction

Finally, cross-camera matches, per-second event logs, and classified stops are combined. For each reconstructed passenger trajectory, the system identifies a boarding stop and an alighting stop, incrementing the corresponding cell in the OD matrix. This process yields detailed flows between official stops, while also capturing demand at illegal roadside pickups and drop-offs. The OD matrices produced in this way serve as the core output of the baseline system, supporting equity analysis and transit planning.

### 2.7 Hybrid detection with head-only fallback

In practice, the baseline detector struggles most under extreme crowding at the exit door, where only passenger heads are visible inside the ROI. To mitigate this, we introduce a hybrid detection strategy that uses a dedicated head detector when local density is high. Concretely, we first run the full-body YOLOv12 model and count the number of tracked boxes whose centers fall inside the door ROI. If this number exceeds a threshold of five people, we treat the scene as overcrowded and switch to a head-only detector trained on CrowdHuman-style annotations. In these frames, IN/OUT events are derived from the head detections rather than the full-body boxes. When the number of people inside the ROI falls back below the threshold, the system returns to full-body detection. This hybrid scheme preserves the simplicity of the baseline while providing a fallback that is better aligned with the visible signal under peak loads.

### 2.8 Door-state aware counting

We also refine ROI-based counting by conditioning it explicitly on door status. In this improved configuration, a door-status signal derived from vehicle telemetry and/or visual cues is used to filter candidate events: ROI crossings at the exit door (Cam-B) are only logged when the door is physically open. This prevents false counts from passengers who stand or move near the exit while the bus is in motion, and from transient occluders such as conductors leaning into the ROI. As we show in Sec. 3, this door-state aware counting substantially improves exit-door accuracy and reduces counting error without changing the underlying detector or tracker.

## 3 Findings So Far

We evaluate the baseline and its variants on 11 manually annotated CCTV segments from Nairobi and Kigali buses, each 3- 8 minutes long, covering a mix of low-density, peak-hour, colour, and monochrome scenes at both doors. For each segment, we annotate per-door entry and exit counts and derive OD matrices by pairing boarding and alighting stops. We report accuracy (fraction of correctly counted passengers), mean absolute error (MAE, in passengers), and the number of complete misses (segments where a method records zero events while the ground truth is non-zero).

Across all segments, the baseline system achieves high entry performance but struggles at the exit door. Aggregated over the 11 videos, entry accuracy is $88.9\%$ with entry MAE of $0.6$ passengers, while exit accuracy drops to $57.6\%$ with exit MAE of $0.6$ passengers (Fig. 1). Total accuracy over both doors is $74.2\%$. These aggregate results are consistent with our earlier condition-wise analysis: in low-density, well-lit footage the baseline counts 40 of 42 passengers correctly (recall $\approx 95\%$) with 4 false positives, and OD matrices match manual tallies, whereas in overcrowded peak-hour boarding at the exit door we observe undercounts of roughly $40\%$, and in monochrome footage we see an average degradation of about $17\%$ points in recall relative to comparable color segments. Posture variation and non-standard door usage further fragment tracklets and introduce ambiguity in directionality, leading to incomplete trip reconstruction even when detection and timestamping remain robust. More details on the data are in Appendix A.1

We then compare three variants to the baseline: (i) `hybrid_det`, which introduces the head-only detector fallback when more than five people occupy the exit-door ROI; (ii) `door_state`, which gates ROI crossings by the door-status signal; and (iii) `all_together (no id repair)`, which combines the improved components without applying cross-camera identity repair. Entry counts are already near ceiling, and all methods achieve similar entry performance: both `door_state` and `all_together` reach $98.0\%$ entry accuracy with MAE $0.5$ passengers, while the baseline and `hybrid_det` remain at $88.9\%$ with MAE $0.6$. The main gains appear at the exit door: compared to the baseline's $57.6\%$ exit accuracy, `all_together` improves to $78.0\%$ and the door-state aware variant further to $82.6\%$, reducing exit MAE from $0.6$ to $0.4$ passengers. In contrast, the `hybrid_det` configuration does not improve exit accuracy over the baseline on these segments. Overall, total accuracy rises from $74.2\%$ (baseline) to $87.8\%$ for `all_together` and $92.4\%$ for the door-state aware configuration. Per-video results are presented in Appendix A.3.

These results suggest that simple, domain-informed cues, especially door-state gating of ROI events and more flexible stop classification that accounts for near-zero motion with open doors, can substantially mitigate the worst failure modes of the baseline at the exit door, even without introducing more complex Re-ID architectures or identity-repair modules. At the same time, the remaining exit errors under extreme crowding and modality shifts indicate that more robust embeddings and trajectory-aware repair will be necessary for reliable OD inference at scale.

## 4 Challenges and Discussion

The above findings reveal several deployment-specific challenges in sub-Saharan bus environments. **Overcrowding and occlusion** at doorways frequently obscure passengers, fragmenting tracklets and driving undercounts of up to $40\%$. **Posture changes** (standing, sitting, leaning) disrupt appearance-based embeddings, leading to broken Re-ID continuity. **Modality shifts**, where cameras switch between color and monochrome, destabilize feature embeddings and degrade OD accuracy by roughly $17$ percentage points. **Visual similarity** among riders in school uniforms or similar clothing causes identity switches across tracklets. Finally, **non-standard door usage**, with boarding and alighting at both front and exit doors, introduces ambiguity into ROI-based event classification.

Our ablations indicate that lightweight, domain-informed cues can substantially reduce these errors. Door-state-aware counting and a more flexible stop definition that includes near-zero motion with open doors improve exit accuracy from $57.6\%$ (baseline) to $82.6\%$ and total accuracy from $74.2\%$ to $92.4\%$, without changing the detector or Re-ID backbone. In contrast, a hybrid head-only detector provides limited benefit on our current clips, suggesting that better use of temporal and door context is more valuable than simply adding another detector. Nevertheless, remaining failures under extreme crowding and modality shifts indicate that more robust embeddings and trajectory-aware identity repair will be needed for reliable OD inference at scale, and that deploying methods such

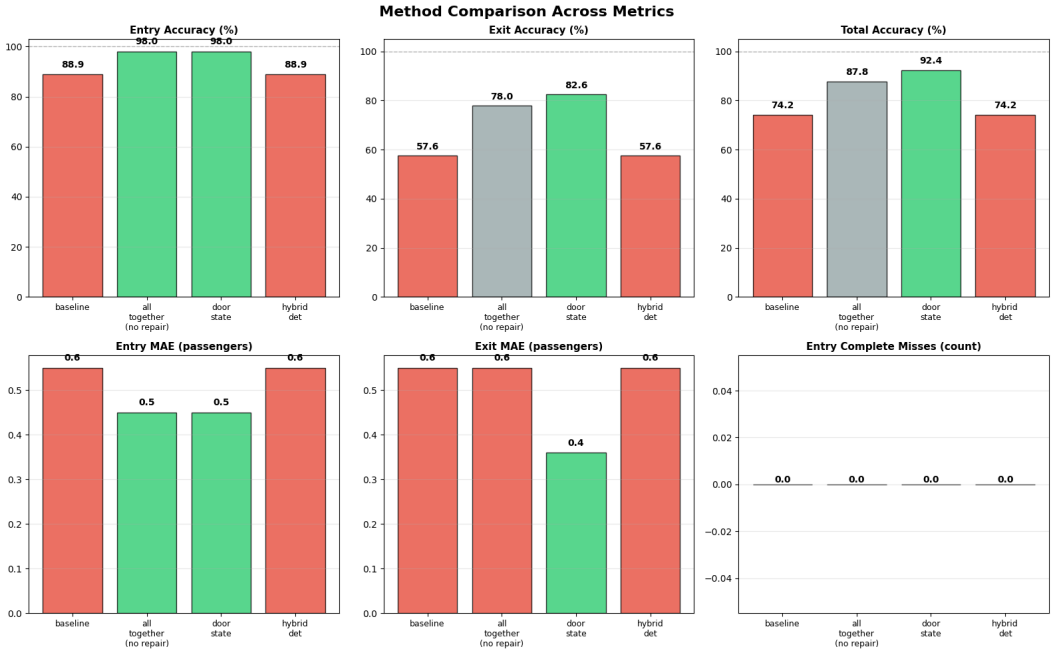


Figure 1: Method comparison across 11 annotated segments. We report entry, exit, and total accuracy (top row) and mean absolute error (MAE, in passengers; bottom row) for the baseline pipeline, the door-state aware variant (`door_state`), the hybrid head-only detector configuration (`hybrid_det`), and an `all_together (no id repair)` variant that combines improved components without identity repair. Entry performance is high for all methods, while exit accuracy and exit MAE improve substantially when incorporating door-state gating and related telematics cues.

as TransitReID in this low-resolution, color/monochrome, roadside-stop setting remains an open challenge.

**Data sharing and reproducibility.** Work with in-vehicle CCTV necessarily involves agreements with operating companies and start-ups that prioritize passenger safety and privacy. In our case, these agreements prohibit releasing raw video or models that could be repurposed for surveillance beyond the scope of this study. Similar constraints apply to other recent transit Re-ID efforts, which often operate on non-public CCTV datasets. As a result, we currently make public only aggregate OD matrices, per-condition metrics, and implementation details of our baseline pipeline. In future work, we plan to explore privacy-preserving benchmarks (e.g., synthetic data or heavily anonymised clips) that still allow meaningful comparisons with this baseline.

## 5 Conclusion

We presented a baseline CCTV pipeline for passenger detection, tracking, OCR timestamping, and ROI-based counting fused with telematics data in SSA buses. Using 3–8 minute annotated clips from Nairobi and Kigali, we showed that the system is feasible under light to moderate passenger loads, achieving high entry accuracy and OD matrices that match manual tallies, but is hindered by overcrowding, occlusion, posture changes, modality shifts, and non-standard door usage. By treating both doors as bidirectional and incorporating door-state aware counting and flexible stop definitions, we align the design more closely with real-world operations and substantially improve exit-door accuracy. These early results highlight both the promise and the limitations of CCTV-based OD inference in SSA transit, and point toward the need for more robust, deployment-focused Re-ID methods.

## 6 Next Steps

To address the remaining failure modes, we plan several extensions. First, we will develop a stronger cross-camera cascade with motion-informed gating and cost-optimal assignment to reduce mismatches under crowding and non-standard door usage. Second, we will incorporate dual-branch embeddings (head and body crops) with grayscale normalization to better handle partial visibility and color/monochrome shifts. Third, we will conduct controlled robustness studies that vary posture and crowd density to quantify their impact on ID stability. Finally, we will extend deployments to longer time windows, enabling analysis of equity impacts, for example, identifying which neighborhoods benefit most from electrified routes and which depend heavily on unscheduled roadside stops.

## References


A. B. R. González, J. J. V. Díaz, and M. R. Wilby. Detailed origin-destination matrices of bus passengers using radio frequency identification. *IEEE Intelligent Transportation Systems Magazine*, 14(1):141–152, 2020.

K. Huang, T. Azfar, J. Reilly, and R. Ke. Transitreid: Transit od data collection with occlusion-resistant dynamic passenger re-identification. *arXiv preprint arXiv:2504.11500*, 2025.

M. Jafari Kang, S. Ataeian, and S. M. Amiripour. A procedure for public transit od matrix generation using smart card transaction data. *Public Transport*, 13(1):81–100, 2021.

C. Kong, T. Guo, and L. He. Research on od estimation of public transit passenger flow based on multi-source data. In *International Conference on Green Intelligent Transportation System and Safety*, pages 589–603. Springer, 2021.

A. N. Larijani, A.-M. Olteanu-Raimond, J. Perret, M. Brédif, and C. Ziemlicki. Investigating the mobile phone data to estimate the origin destination flow and analysis; case study: Paris region. *Transportation Research Procedia*, 6:64–78, 2015.

X. Liu, P. Van Hentenryck, and X. Zhao. Optimization models for estimating transit network origin–destination flows with big transit data. *Journal of Big Data Analytics in Transportation*, 3(3):247–262, 2021.

J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 542–551, 2019.

K. Ozbay, N. Shlayan, H. Nassif, et al. Real-time estimation of transit od patterns and delays using low cost-ubiquitous advanced technologies. 2017.

Y. Shimada, M. Takagi, and Y. Taniguchi. Person re-identification for estimating bus passenger flow. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 169–174. IEEE, 2019.

A. Zalewski, D. Sonenklar, A. Cohen, J. Kressner, and G. Macfarlane. Public transit rider origin-destination survey methods and technologies. Technical report, TCRP Synthesis, No. 138, Washington, DC, 2019. URL `https://doi.org/10.17226/25428`.

J. Zhao, C. Li, Z. Xu, L. Jiao, Z. Zhao, and Z. Wang. Detection of passenger flow on and off buses based on video images and yolo algorithm. *Multimedia Tools Appl.*, 81(4):4669–4692, Feb. 2022. ISSN 1380-7501. doi: 10.1007/s11042-021-10747-w. URL `https://doi.org/10.1007/s11042-021-10747-w`.


## A Supplementary Material

We include data and annotation details, the full baseline pipeline diagram, and per-clip counting results to support reproducibility and interpretation of the main findings.

### A.1 Data and annotation details

In both cities, bus configurations differ by door count, camera count and placement. **Kigali** buses generally have two doors, an entrance at the front and an exit in the middle of the cabin. Cameras: Driver Cabin ($Cam\text{-}Dr_k$), placed directly above the driver's cabin and capturing only the driver; Dashboard ($Cam\text{-}Dash_k$), facing outward towards oncoming traffic; Front Door ($Cam\text{-}A_k$) positioned at the front of the bus and faces down the aisle; captures boardings, front-door alightings and initial movements on the aisle; Exit Door ($Cam\text{-}B_k$) located opposite the exit door, facing the doorway; captures alightings and occasional rear-door boardings; and where present Rear ($Cam\text{-}Re_k$) placed at the rear and facing forward for extended aisle trajectories. In contrast,

**Nairobi** buses have a single front door that serves as both the entrance and the exit. Like the buses in Kigali, these buses are equipped with Driver Cabin Cameras ($Cam\text{-}Dr_n$) and Dashboard Cameras ($Cam\text{-}Dash_n$). Additionally, each bus has a Front Door Camera ($Cam\text{-}A_n$) located near the front facing the door, capturing both boarding and alighting.

### A.1.1 Recording & telematics.

In both cities, CCTV footage runs from $\sim$05:00 to $\sim$23:00. Videos are $\approx$30–45 minutes long with on-frame timestamps. Telematics provide GPS (latitude, longitude) and vehicle signals (wheel-based speed, odometer, state-of-charge (SoC), voltage/current, and energy consumed/recuperated/idle/charged/used), each time-stamped with a unique telematics ID.

### A.1.2 Clip selection for tuning/validation.

From $Cam\text{-}A_k$, $Cam\text{-}B_k$, and $Cam\text{-}A_n$, we extract 3–4 minute segments per interval under the following criteria: (i) uncrowded, (ii) medium-high crowding, (iii) partial occlusions from bus hardware, (iv) color-only or monochrome-only, and (v) alternating monochrome$\leftrightarrow$color (e.g., IR night vision). For Kigali, paired $Cam\text{-}A_k$/$Cam\text{-}B_k$ views ensure temporal continuity for cross-view ReID. Two highly characteristic videos are densely annotated using Computer Vision Annotation Tool (CVAT) for held-out testing; the final pipeline is evaluated on footage from both cities. Images extracted from the video form part of our Indoor data that will be used to fine-tune the ReID model.

### A.1.3 Data splits.

Because routes repeat daily as terminal-to-terminal *round trips*, we use *day-level, temporally disjoint* splits to prevent temporal/scene leakage while preserving deployment realism. For **Kigali**, we use the E-bus launch days *March 21* (15:00–22:00) and *March 26* (04:00–22:00) for training/validation, and *March 31* (04:00–22:00) for testing. When **Nairobi** data is included, we mirror this policy with day-level disjoint train+val and test days.Trip/route segmentation uses wheel speed and GPS to partition days into round trips and per-stop dwell intervals; there is no overlap in timestamps, telematics IDs, or trip IDs across splits. We stratify evaluation by time-of-day bins aligned to demand: AM peak (07:00–09:00), Midday (11:00–12:00), PM peak (17:00–19:00), and Off-peak (14:00–16:00), as well as by lighting (day/night), stream (color/monochrome), and door-ROI occupancy bins $[0, 5)$, $[5, +)$ persons/frame (centers within $\mathcal{P}_c$ or a thin band around $\ell_c$).

## A.2 Baseline pipeline diagram

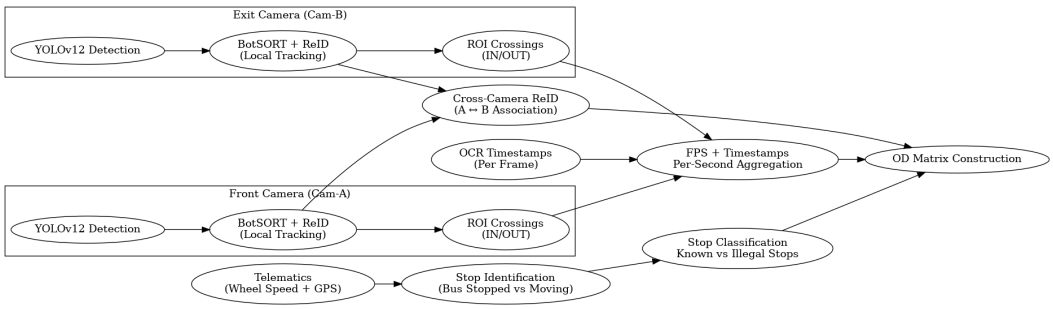


Figure 2: Baseline two-stream pipeline for OD inference. Each camera (Cam-A at the front, Cam-B at the exit) runs YOLOv12 detection with BotSORT+OSNet tracking to produce local tracklets and ROI-based IN/OUT events. OCR-extracted timestamps and FPS align events to a per-second timeline, while telematics (wheel speed and GPS) identify official and illegal stops. These signals are fused to construct stop-level OD matrices; later sections add hybrid head-only detection and door-state aware counting on top of this baseline.

## A.3 Per-clip counting results

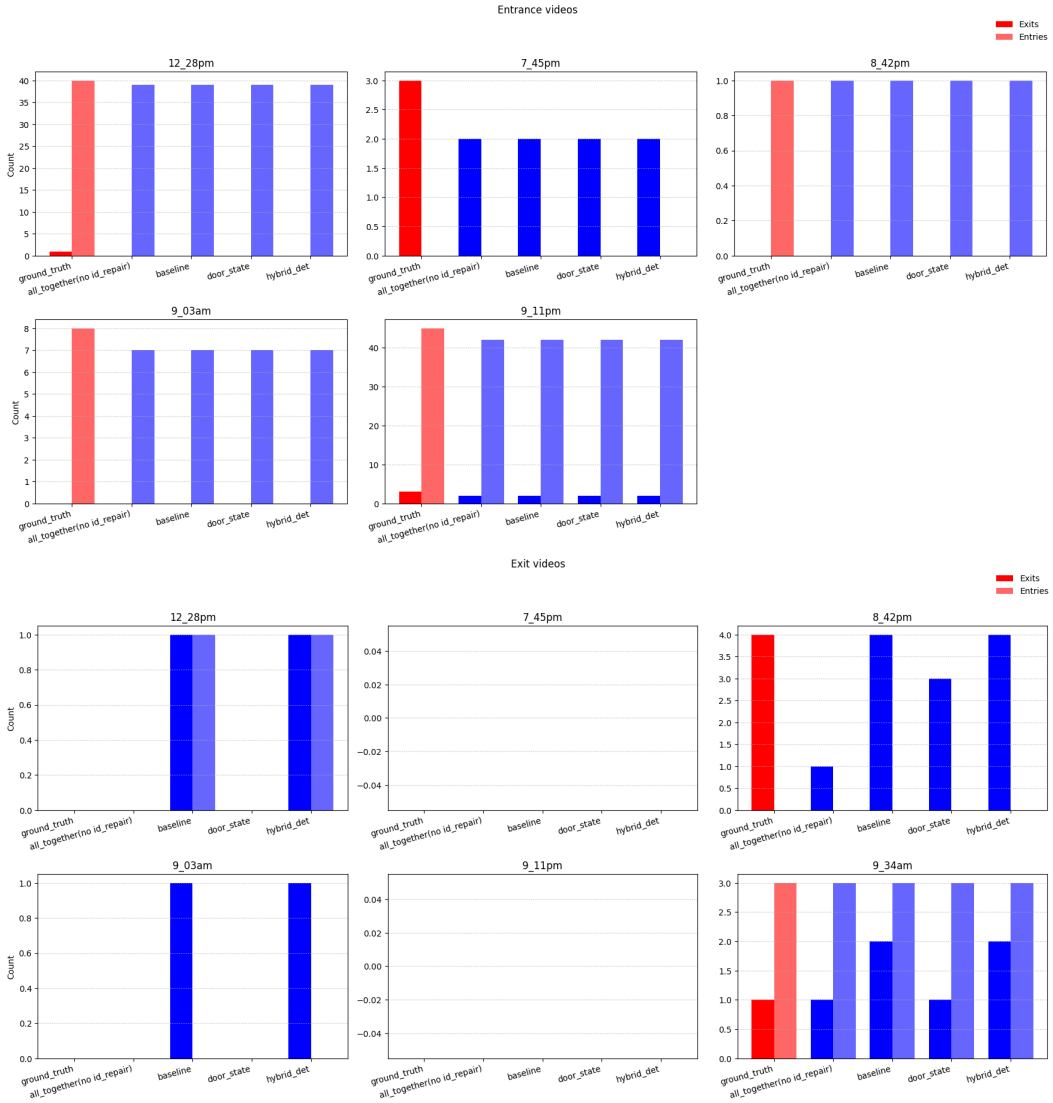


Figure 3: Per-clip entry (top) and exit (bottom) counts across 11 annotated segments. Red bars show ground-truth counts; blue bars show the baseline, `door_state`, and `hybrid_det` variants. Blank panels (no visible bars) correspond to clips in which no events occurred during the 3–8 minute window, so all methods correctly report zero entries or exits.