DOORFUSE: Stop-Level Boarding/Alighting Counting from Bus CCTV and Telematics

Anonymous Author(s)

Abstract

Vision-based passenger counting on buses struggles with doorway-specific challenges: occlusion, track fragmentation, and persistent non-passengers generating spurious counts. We present DOORFUSE, a door-aware system fusing five signals—trajectory, orientation, motion, biometric, and consistency—to resolve ambiguous crossings. Key mechanisms include door-state gating, head-assisted tracking for occlusion robustness, pose-based directional scoring, and behavioral filtering of staff and touts. Evaluated on 4.2 hours of bus CCTV from Nairobi and Kigali, DOORFUSE reduces MAE by 78–85% and improves F1 from 0.66–0.77 to 0.93–0.94 versus a tracking baseline. Ablations confirm each component addresses distinct failure modes, with head-assisted tracking providing +34 percentage points exit F1.

CCS Concepts

• **Applied computing** → **Transportation**; • **Computing methodologies** → **Tracking**; *Activity recognition and understanding*; *Object detection.*

Keywords

automatic passenger counting, transit systems, video analytics, edge deployment, urban mobility

ACM Reference Format:

Anonymous Author(s). 2018. DOORFUSE: Stop-Level Boarding/Alighting Counting from Bus CCTV and Telematics. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Rapid urbanization in sub-Saharan Africa (SSA) has intensified pressure on bus systems to deliver reliable and efficient service [2]. These environments are characterized as rapidly growing cities with limited vehicle ownership, heavy transit use, and few if any other public transit options beyond buses. However, many transit agencies operating in these settings face constraints such as limited resources, aging infrastructure, and a lack of high-resolution demand data, which hinders the identification of peak loads, route bottlenecks, and persistent service gaps. Measurement of passenger demand at fine temporal and spatial granularity offers a practical approach to inform scheduling, fleet allocation, and service-quality monitoring.

Passenger counting serves as a fundamental measurement for public transit planning. Accurate boarding and alighting counts enable capacity management, mitigate overcrowding risks, and assist agencies in prioritizing interventions and evaluating their impact [9]. In addition to aggregate ridership, stop-level demand profiles can identify underserved areas and routes, and differentiate official stops from frequent informal roadside pickups, which are prevalent in SSA operations but are inadequately captured by coarse reporting. These insights facilitate cost-effective resource management and an enhanced passenger experience, enabling these vital services to scale efficiently under dynamic conditions.

Prior work has measured passenger demand using manual audits or automated systems based on fare transactions or onboard sensing [14]. Automated Fare Collection (AFC) can infer origin–destination (OD) flows are derived from smart-card data but often rely on behavioral assumptions (e.g., round trips) [4]. Dedicated sensing systems, such as Automated Passenger Counters (APCs), improve automation [8], but installation costs, maintenance overhead, and limited coverage can complicate large-scale deployment – especially across heterogeneous fleets.

Computer vision offers a promising alternative, as it can leverage existing CCTV streams to automate detection, tracking, and re-identification. Nevertheless, deploying such pipelines on SSA buses presents several unique challenges: heavy occlusion and crowding at doorways; heterogeneous camera viewpoints and door layouts across vehicles; frequent appearance changes due to lighting and camera mode shifts, such as color, monochrome, or infrared switching; and significant near-door motion during boarding, including queues at thresholds and persistent non-passengers such as touts (staff who serve roles as hawkers and fare collectors, also known as conductors) or door loiterers, which can result in spurious or duplicate counts. Furthermore, many current approaches provide only aggregate counts and do not reliably associate events with specific stops, thereby limiting downstream planning and behavioral analysis.

In our target cities, the transition to electric fleets presents a timely opportunity, as buses are increasingly equipped with pre-installed cameras and time-stamped telematics streams. This study involves collaboration with an electric bus provider in Nairobi, Kenya, and Kigali, Rwanda. The buses in the provider's fleets have a variety of designs, varying in door layouts and camera placements. CCTV footage, combined with telematics signals such as GPS and wheel-based speed, enables a multimodal (video and telematics) pipeline that links door events to vehicle motion and location for stop-level analysis.

Proposed system and evaluation overview. We propose DOORFUSE, a door-aware **bus-CCTV + telematics** passenger counting system for stop-level boarding/alighting measurement under real

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

SSA operating conditions. Starting from a baseline vision pipeline (person detection, tracking, door-ROI line-crossing, and OCR timestamp parsing), our core contribution is a **weighted multi-signal counter** that resolves ambiguous door events by fusing complementary evidence:

(1) **Door-state gated counting** to suppress spurious crossings when doors are closed;
(2) **Crowd-aware dual-detector cues** (full-body + head) to stabilize tracklets under occlusion and dense queues;
(3) **Weighted decision fusion** aggregating five signals – orientation, motion, trajectory, biometric, and consistency – to produce robust IN/OUT decisions that reduce errors from identity fragmentation, brief occlusions, and crowd-induced jitter;
(4) **Role filtering** using persistent behavioral patterns to remove non-passenger actors that otherwise generate systematic false counts.

Evaluation on 24 video clips (comprising a total of 4.2 hours of footage) from Kigali and Nairobi demonstrates substantial improvements over a strong tracking baseline. DOORFUSE reduces entry MAE by 78% (5.56 → 1.20) and exit MAE by 85% (6.88 → 1.04), while improving F1 score from 0.66–0.77 to 0.93–0.94 across multiple sets. Ablation studies isolate each mechanism's contribution: head-assisted tracking provides the largest single improvement (+34% exit F1), followed by pose-based scoring (+15% entry F1) and staff/tout filtering (93% reduction in entry overcounting).

Contributions.

(1) A door-aware, weighted multi-signal counting pipeline that reduces MAE by 78–85% and improves F1 by 17–27 percentage points over state-of-the-art tracking using CCTV on heterogeneous bus designs in SSA;
(2) Multimodal alignment of visual events to stop-level context via OCR timestamps and vehicle telematics, enabling stop-type classification (official vs. informal); and
(3) A two-city evaluation with ablations quantifying module-level gains: head-assisted tracking (+34 pp exit F1), pose-based scoring (+15 pp entry F1), staff/tout filtering (93% overcounting reduction), and door-state gating (44% MAE reduction).

2 Related Works

This section reviews passenger-counting and origin– destination (OD) inference methods, focusing on three primary constraints: (i) the requirement for stop-level boarding and alighting measurement, (ii) the practical trade-offs among coverage, cost, and detail, and (iii) the doorway as the principal source of ambiguity in actual bus interiors.

2.1 Ticketing and sensor-based measurement

Transit agencies commonly estimate ridership and, in some cases, OD structure using ticketing logs and onboard sensors. These approaches involve trade-offs among coverage, cost, and behavioral characteristics. Manual tallies are straightforward but intermittent and prone to error at scale. Door-mounted sensing systems, such as Automated Passenger Counters (APCs), provide reliable boarding and alighting totals on instrumented vehicles. However, installation and maintenance overhead limit their coverage across large fleets, especially where retrofitting is costly [8, 14].

Automated Fare Collection (AFC) and smart-card systems provide scalable measurement of boardings and can facilitate OD inference from tap data. However, these systems often require assumptions about alighting, transfers, or return trips, and may fail when riders do not tap out or when behavior deviates from modeling assumptions [4]. Proxy signals such as RFID, Bluetooth, or mobile-phone traces, along with GIS and ticketing data fusion, can enhance spatial coverage and recover coarse OD patterns. These methods, however, may be sensitive to device ownership, sampling bias, and data access constraints [3, 5, 6, 11].

These limitations motivate using existing onboard CCTV, now increasingly available with vehicle telematics, to directly observe doorway events and align them with stops. This approach enables finer-grained stop-level boarding and alighting profiles without requiring new per-vehicle sensing infrastructure.

2.2 Vision-based passenger counting

Vision-based passenger counting recovers boardings and alightings directly from video by detecting individuals and analyzing motion through a doorway or region of interest (ROI). Earlier studies used handcrafted motion cues and background modeling to address challenges like changing lighting and occlusion. Recent research employs deep detectors and trackers, which improve robustness to clutter but still face significant challenges in the doorway region. Severe partial occlusion, close interactions, and near-door loitering can cause missed detections, fragmented tracks, and duplicate counts, especially in fixed-view bus cameras with limited resolution and compression artifacts.

2.3 Detection, multi-object tracking, and ReID

Modern pipelines typically decompose the problem into three stages: detection, within-camera multi-object tracking (MOT), and optional cross-camera association via re-identification (ReID). Recent MOT systems enhance association by integrating motion models with appearance cues. Strong baseline tracking-by-detection approaches from the SORT and ByteTrack families show that careful association design can yield substantial improvements without end-to-end training [1, 15]. However, identity stability deteriorates in crowded, low-quality footage where detections fragment and appearance cues become unreliable. For ReID, compact embedding models such as OSNet enable efficient appearance matching [16]. Cross-view and cross-modality shifts, such as color to monochrome or infrared, and heavy occlusion can distort embeddings and increase mismatches unless the pipeline explicitly addresses these shifts.

2.4 Crowd-aware cues under occlusion

In crowded transit scenes, the most informative visual evidence changes with passenger density. When bodies are heavily occluded near the threshold, head cues often remain visible and provide a more reliable signal than full-body bounding boxes. Crowd-focused datasets and detectors emphasize this scenario and motivate hybrid strategies that revert to head-based cues during high-density door intervals [13]. However, switching between signals introduces new

failure modes, such as double counting, inconsistent geometry, and tracker disruption, unless the pipeline carefully manages fallback activation and event logging.

2.5 Context and multimodal alignment

Another area of related work uses contextual information and additional sensors to reduce ambiguity. In transit settings, video events are linked to vehicle state (stopped or moving), door state (open or closed), and location (at a stop or mid-block). Aligning video timestamps with telematics data can suppress entire classes of false events, such as door crossings while the vehicle is moving, and supports downstream applications including stop-level counts and OD matrices. Less commonly, door state is treated as a primary control signal for counting and for conditioning association behavior and identity-repair policies during periods of turbulence at the threshold when doors are open.

2.6 Positioning of this work

This work addresses reliable stop-level boarding and alighting measurement under real SSA bus conditions, with frequent doorway churn, occlusion, camera heterogeneity, and appearance modality shifts. Rather than introducing a new detector or tracker in isolation, the study examines a door-aware CCTV and telematics pipeline, systematically ablating targeted doorway mechanisms while maintaining the remaining stages. Beyond door-state gating and crowd-aware body and head cues, the proposed counter resolves ambiguous door events through a weighted multi-signal fusion of orientation, motion, trajectory, and biometric evidence, and evaluates the contribution of these mechanisms to counting reliability across two cities.

2.7 Vision–language models

Recent advances in vision–language pretraining enable open vocabulary recognition and text-conditioned localization, as demonstrated by CLIP-style representations and open-vocabulary detectors such as OWL-ViT and Grounding DINO [7, 10, 12]. Although promising, these models are not suitable as direct replacements for door-line counting, which requires stable frame-to-frame identities and consistent online performance under compression artifacts and infrared or monochrome shifts. In this context, vision–language models are best used as offline tools to accelerate labeling and auditing, such as proposing candidate door-open windows or flagging likely loitering and occlusion segments, rather than serving as the primary online perception system.

3 Problem Setting and Design Requirements

This section defines the deployment environment, operational constraints, and the doorway-centric failure modes that motivate DOORFUSE's design.

3.1 Deployment setting

This study targets passenger counting on urban buses in Nairobi (Kenya) and Kigali (Rwanda) under uncontrolled operational conditions. Deployments vary across vehicles due to differences in door layout, interior configuration, and camera placement.

Kigali buses typically have two doors: a front entrance and a mid-bus exit. Camera viewpoints include a front-door camera oriented down the aisle, capturing boardings and initial movement, and an exit-door camera facing the rear doorway. Nairobi buses use a single front door for both boarding and alighting, with a front-door camera positioned to observe bidirectional doorway traffic. In both cities, camera mounts vary in height and angle, producing cabin-facing and doorway-facing views with heterogeneous fields of view.

3.2 Doorway failure phenomenon

Doorway regions represent the primary source of counting errors. Three failure modes dominate: (1) **Closed-door false positives:**passengers queuing near the threshold intersect the ROI, producing spurious crossings; (2) **Open-door fragmentation:**frequent occlusions at steps cause track breaks and identity switches, leading to duplicate counts; and (3) **Occlusion-induced detection failure:**under crowding, full-body detections degrade while heads remain visible, challenging appearance-based re-identification. These effects motivate door-state gating, doorway-specific continuity controls, and head-assisted tracking.

3.3 Operational constraints

Transit-operator requirements define practical feasibility. The system must handle frequent overcrowding and turbulent near-door motion, non-standard door usage (e.g., mixed flows or reversed entry/exit), and privacy constraints that preclude reliance on identity recognition. Deployment is constrained to existing CCTV infrastructure without per-vehicle retrofits, running on edge devices with bounded latency. Video quality is uncontrolled: glare, night-vision artifacts, missing frames, compression degradation, and recorder overlays are common. The system must maintain performance despite these conditions and produce audit-ready records.

3.4 Required outputs

The system produces: (i) a time-stamped event stream of doorway crossings with direction (IN/OUT) and confidence scores, (ii) stop-level boarding and alighting counts aligned to GPS-derived stop windows, (iii) trip-level totals by aggregation, and (iv) per-event diagnostic metadata (gating decisions, signal contributions) for auditability. Counts must exclude persistent non-passenger actors such as staff and touts.

3.5 Design requirements

R1: Doorway accuracy under crowding. Counting must remain stable under heavy occlusion, brief interactions, and oscillatory motion near the counting line. The system should minimize both missed detections from occlusion and duplicate counts from track fragmentation.

R2: Non-passenger filtering. Conductors collecting fares, touts lingering at the entrance, and loiterers near full buses must not trigger repeated counts. The system must distinguish these actors from genuine passengers without suppressing valid events.

R3: Stop-level alignment. Crossing events must be reliably attributed to stops, accommodating both scheduled stops and informal

Table 1: Dataset summary by city and bus type.

City	Bus	Clips	Duration	View
Kigali	Type 1	5 paired	3–8 min per	Ent + Exit
Kigali	Type 2	4 paired	15 min per	Ent + Exit
Nairobi	Type 1	4	16 min per	Combined
Nairobi	Type 2	2	6 + 25 min	Combined
Total		**24**	**~4.2 hrs**	

roadside pickups. Alignment must tolerate noisy GPS signals, brief stop-and-go motion, and minor timestamp offsets.

R4: Low operational overhead. The pipeline should use existing CCTV and commonly available telematics, avoid per-fleet retraining, and maintain computational efficiency through lightweight mechanisms.

R5: Auditability. Outputs must include sufficient metadata to reproduce decisions and diagnose failures, including per-event confidence, gating state, and dominant signal contributions.

4 Data and Evaluation Protocol

DOORFUSE is evaluated on operational bus CCTV footage from four different bus designs, two each found in Nairobi and Kigali, with corresponding vehicle telematics. Evaluation is limited to entrance/exit-view streams (Section 3.1), as these enable direct measurement of boarding and alighting.

4.1 Data acquisition

CCTV footage is recorded during routine operations (~05:00–23:00) with on-frame timestamps extracted for second-level temporal alignment. Video quality is uncontrolled and reflects operator-grade equipment: lighting variation, IR/monochrome night modes, motion blur, compression artifacts, and frame drops. Telematics (GPS, wheel speed, odometer) are logged over matching intervals and used to segment stop windows for stop-level reporting.

4.2 Dataset scope

Table 1 summarizes the evaluation set. Kigali provides *paired* entrance/exit door views (i.e., those buses have two doors); Nairobi provides a single front-door view capturing both boarding and alighting (i.e., those buses have only one door). Stop-level ground truth is obtained via third-party manual annotation on development and held-out test clips.

4.3 Clip selection

Clips are stratified to emphasize failure-prone conditions rather than clean boarding scenes: heterogeneous door layouts, peak and off-peak loads, night-time modality shifts (color to IR/monochrome), and high-churn doorway behaviors including closed-door queueing, near-threshold loitering, and repeated staff or tout crossings. This ensures evaluation targets DOORFUSE's core challenges: threshold occlusion, spurious line crossings, and persistent non-passenger actors.

For ablation tests to isolate individual component contributions, clips are grouped into five overlapping evaluation sets based on conditions each module targets, as described in Table 2.

Table 2: Ablation evaluation sets. B/A = boarding/alighting counts. Sets overlap; durations are per-set totals.

Set	Target Condition	Clips	Dur.	B/A
S1	Door-state gating	7	48 min	137/68
S2	Crowd density (dual det.)	6	52 min	178/26
S3	Staff/tout/line-seating	8	67 min	157/37
S4	Bidirectional traffic	7	71 min	120/110
S5	Modality robustness	8	58 min	134/76

4.4 Data partitioning

Partitions are disjoint at the *clip level* to prevent leakage: **(i) Training and Validation sets**: labeled frames for detection, door-state, and ReID fine-tuning, split by clip rather than frame; **(ii) Dev set**: separate clips with third-party ground truth for threshold and fusion-weight selection; **(iii) Test set**: held-out clips from different recording intervals, used once for final reporting. Results are stratified by city, bus type, lighting regime, and crowding level to expose failure modes beyond aggregate accuracy.

To ensure the privacy of bus operators and passengers, access to CCTV data used in the study is restricted to approved researchers with controlled permission. Results are reported as aggregate counts; figures use face obfuscation and tight cropping where appropriate.

5 System Overview

DOORFUSE is a multi-stage verification architecture that converts bus CCTV footage into stop-level passenger counts. This section describes the design rationale; implementation specifics follow in Section 6.

5.1 Architecture overview

The pipeline addresses requirements R1–R5 through five stages (Figure 1):

(1) **Digital Synchronization** aligns video frames with vehicle telematics using OCR-extracted timestamps, enabling stop-level attribution (R3).
(2) **Trajectory Formation** detects individuals and links detections across frames using motion and appearance cues, producing positional histories for behavioral analysis.
(3) **Door-Eligible Gating** restricts candidate generation to intervals when the door is open, suppressing closed-door false positives from queuing passengers (R1).
(4) **Multi-Signal Fusion** scores each candidate using complementary behavioral evidence, distinguishing committed crossings from oscillatory or ambiguous interactions (R1, R2).
(5) **Stop Aggregation** groups verified events into telematics-derived windows, distinguishing official stops from informal roadside exchanges (R3).

Each stage addresses specific failure modes identified in deployment. We describe the design rationale for key components below.

5.2 Head-assisted tracking

Standard full-body tracking fails predictably in crowded doorways: as passengers cluster at steps and thresholds, mutual occlusion causes detection dropout and track fragmentation. However, we

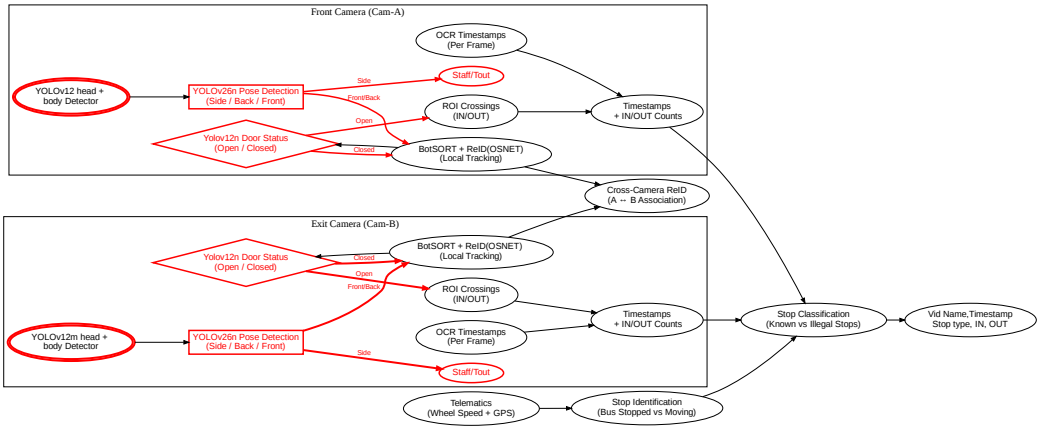

Figure 1: Two-stream per-door pipeline. Black blocks show the baseline: YOLOv12 detection, BoT-SORT+ReID tracking, and ROI line-crossing. Non-black blocks show DOORFUSE extensions: head+body fusion for occlusion handling, pose-based orientation for staff/tout filtering, door-state gating, and timestamp OCR for audit logs. Telematics (wheel speed + GPS) segments stops and classifies known vs. informal stops.

observed that heads typically remain visible even when bodies are fully occluded; passengers naturally maintain head separation to see their path.

DOORFUSE exploits this asymmetry through a Ghost Manager that preserves track identity when bodies temporarily disappear. When a tracked body vanishes while a spatially consistent head persists, the system injects a synthetic body detection anchored to the head's position. This maintains the track's positional history during occlusion, preventing duplicate counts that occur when a single passenger is assigned multiple track IDs.

The key design choice is *conservative injection*: synthetic bodies are assigned lower confidence than real detections, ensuring the tracker prefers genuine observations when available. Injection is also spatially restricted to the near-door region where identity preservation matters most for counting accuracy.

5.3 Door-eligible gating

Analysis of false positives showed that a significant portion stems from periods when doors are closed: passengers lining up outside the vehicle, waiting at the entrance to board, or exceeding bus capacity, which leads to congestion near the doorways as surplus passengers remain by the doors and repeatedly intersect with the detection boundary during transit. Such interactions with the doorway region of interest cannot be geometrically differentiated from actual crossings.

DOORFUSE addresses this by monitoring door state and suppressing candidate generation when the door is classified as closed. Two refinements improve robustness:

Temporal smoothing. Door detections exhibit frame-to-frame noise from partial visibility and motion blur. An exponential moving average filter prevents rapid state oscillation that would repeatedly enable and disable counting.

Crossing grace period. Passengers mid-transit when the door closes should still be counted. Tracks that enter the doorway polygon during a door-open interval remain count-eligible for a brief grace window to accommodate the physical time required to complete a crossing.

5.4 Multi-signal fusion scoring

Simple line-crossing triggers produce unacceptable error rates in crowded doorways. Passengers pause at thresholds, oscillate under crowding pressure, and step partially in and out while waiting. Staff and touts cross repeatedly. A single geometric criterion cannot distinguish these cases.

DOORFUSE treats each crossing as an evidence aggregation problem. Rather than requiring any single definitive signal, the system combines multiple weak indicators into a composite confidence score:

$$\text{Score}(e) = \sum_i w_i \cdot s_i(e), \tag{1}$$

for all crossings i, where $s_i(e) \in [-1, 1]$ indicates directional evidence (positive for boarding, negative for alighting) and weights w_i reflect signal reliability. Five signals contribute:

- **Trajectory displacement**: How far past the counting line did the person travel? Deeper penetration indicates committed crossing rather than threshold oscillation.
- **Motion velocity**: Is movement toward or away from the bus interior? Sustained directional velocity distinguishes intentional transit from jitter.
- **Pose orientation**: Which way is the person facing? Boarding passengers typically face inward; alighting passengers face outward.
- **Head visibility**: Was the head consistently detected? Head presence anchors identity confidence, particularly when body detection is intermittent.

- **Path consistency**: How straight was the trajectory? Passengers in transit follow relatively direct paths; loiterers and staff exhibit meandering or back-and-forth patterns.

A critical design choice is **crowd-adaptive weighting**. Pose estimation degrades under occlusion; shoulder keypoints become unreliable when passengers overlap. As near-door density increases, DOORFUSE automatically shifts weight from pose-dependent signals toward trajectory-based evidence, which remains measurable even when pose extraction fails.

5.5 Non-passenger filtering

Three actor categories require explicit handling to satisfy Requirement R2 (i.e., to count passengers once and only once):

Staff and conductors occupy the doorway for extended periods, often facing sideways to interact with passengers. Naive counters record each threshold crossing, producing dozens of false counts per trip. DOORFUSE identifies staff through converging behavioral evidence: sustained side-facing orientation, positional anchoring (remaining near an initial location), and pacing patterns (repeated back-and-forth movement along the door axis). Requiring multiple signals prevents false classification of passengers who happen to pause while side-facing.

Touts and vendors exhibit different patterns: they enter and exit the polygon repeatedly, approach passengers at close range, and show active movement rather than stationary waiting. DOORFUSE tracks polygon re-entry frequency, proximity events with other passengers, and movement activity to distinguish touts from passengers waiting to alight.

Loiterers may remain in the doorway zone when the bus is full, unable to board. Unlike touts, they are typically passive, with few re-entries, minimal engagement, and low activity. DOORFUSE avoids penalizing long dwell time alone, instead requiring active behavioral indicators before suppressing a track.

5.6 Stop-level aggregation

Transit agencies require counts attributed to stops, not raw event streams. DOORFUSE segments time into stop windows using GPS position and wheel speed. Windows are classified as *official stops* (near mapped coordinates) or *informal exchanges* (stopped or slow-moving away from mapped stops). This distinction matters operationally: informal pickups are common in our deployment cities and represent legitimate ridership that should not be discarded.

Verified crossing events inherit the active stop window at their timestamp, producing per-stop boarding and alighting totals. When GPS quality degrades, the system falls back to speed-only segmentation, flagging affected windows for potential manual review.

6 Implementation

This section details the engineering realization of DOORFUSE, including model selection, parameter values, and the empirical basis for threshold choices. The system is implemented in Python using PyTorch and processes video as a streaming pipeline with frame-sequential execution. Table 3 summarizes key parameters; justifications follow.

Table 3: Key parameters and values for DOORFUSE.

Component	Parameter	Value
Fusion Scoring	Base threshold	0.40
	High-confidence threshold	0.55
	Minimum evidence	0.25
Staff Detection	Side-orientation ratio	0.75–0.95
	Min observation frames	25
	Anchor distance	80 px
	Displacement override	100 px
Ghost Injection	Association distance	120 px
	Injection confidence	0.6
	Max persistence	90 frames
Motion Validation	Teleport ceiling	250 px
	Rolling multiplier	6.0×
Door Gating	EMA window	5 frames
	Crossing grace period	15 frames

6.1 Detection and tracking stack

- **Person detection:** YOLOv12m with confidence threshold 0.25, deliberately low to avoid missing partially-visible passengers, with downstream filtering handling false positives.
- **Head detection:** A separate YOLOv12m instance detects heads for ghost injection. Detected boxes are filtered by aspect ratio (0.5–1.8) and maximum height (150 pixels) to exclude full-body detections.
- **Pose estimation:** YOLOv26m-pose extracts 17-point keypoints. Shoulders (indices 5, 6) are used for orientation analysis with keypoint confidence threshold 0.4.
- **Tracking:** BoT-SORT combines Kalman filtering with OSNet re-identification embeddings, maintaining identity through brief doorway occlusions.
- **Door state:** YOLOv12-nano monitors the doorway region every 3 frames; intermediate frames inherit the previous state.
- **Timestamp extraction:** Microsoft Florence-2 performs OCR on timestamp overlays at 1Hz for video-telematics alignment. This component enables alignment between video events and telematics streams, which arrive on independent clocks.

6.2 Fusion weights

Signal weights were determined through grid search on the development set, optimizing for balanced precision and recall across boarding and alighting:

$$(w_{\text{ori}}, w_{\text{mot}}, w_{\text{traj}}, w_{\text{bio}}, w_{\text{con}}) = (0.25, 0.15, 0.30, 0.20, 0.10) \quad (2)$$

- **Trajectory (0.30):** Highest weight; geometric displacement is directly observable and robust to occlusion.
- **Orientation (0.25):** Strong directional signal when pose is reliable, but degrades under crowding.
- **Biometric (0.20):** Anchors identity confidence via head detection.
- **Motion (0.15):** Captures velocity direction; moderate weight reflects noise at low frame rates.

Table 4: Signal weights under varying crowd density.

Density	w'_o	w'_m	w'_t	w'_b	w'_c
Empty (d=0)	0.25	0.15	0.30	0.20	0.10
Moderate (d=6)	0.16	0.12	0.37	0.25	0.10
Heavy (d=10)	0.10	0.11	0.42	0.28	0.10

- **Consistency (0.10):** Small bonus for straight paths; partially redundant with staff detection.

Crowd-adaptive weighting. As door-region density d increases, DOORFUSE shifts weight from pose-dependent signals to trajectory-based signals using degradation factor $\gamma(d) = \min(0.6, 0.06d)$. Table 4 shows the effect. This enhances performance in both dense and sparse situations.

6.3 Counting thresholds

Base acceptance (0.40): Candidates require |Score| $\geq$ 0.40. Lower thresholds led to more false positives from oscillating passengers; higher thresholds miss legitimate crossings.
High-confidence fast-pass (0.55): Candidates above 0.55 are counted immediately without full polygon dwell time, corresponding to the 90th percentile of genuine crossing scores.
Minimum evidence (0.25): Candidates below 0.25 are immediately rejected, preventing accumulation of marginal candidates.

6.4 Staff detection parameters

Side-orientation ratio threshold (0.75–0.95): Staff classification requires side-facing orientation in >75% of observed frames, with full confidence at 95%. The shoulder width ratio threshold of 0.28 (relative to bounding box width) distinguishes side-facing (<0.28) from front-facing (>0.28) poses.
Minimum observation frames (25): Approximately 0.8 seconds at 30fps. Shorter windows produced false staff labels on passengers who paused briefly while side-facing; longer windows delayed detection of actual staff.
Anchor distance (80 pixels): Staff typically remain within ~80 pixels of their initial position. This threshold corresponds to approximately 0.5 meters at typical camera distances, reflecting the physical workspace of a tout collecting fares at the door.
Displacement override (100 pixels): Tracks exhibiting side orientation but moving >100 pixels total displacement are reclassified as passengers. This override prevents false staff labels on passengers who board while looking sideways (e.g., speaking to someone outside).
Long-stay threshold (450 frames): Tracks present for >15 seconds without crossing, regardless of orientation ratio, are classified as staff. This catches conductors who face forward while stationary.

6.5 Tout and loiterer parameters

Polygon dwell threshold (300 frames): Approximately 10 seconds at 30fps. This duration distinguishes brief passenger pauses from extended tout presence. The value was set conservatively high to avoid penalizing passengers delayed by crowding.
Re-entry threshold (3): Passengers typically enter the polygon once (boarding) or exit once (alighting). Three or more entries suggest repeated in-out behavior characteristic of touts.
Engagement distance (100 pixels): Close approaches within 100 pixels (~0.6 meters) to other passengers are logged as engagement events. Touts accumulate many such events; waiting passengers few.
Activity ratio threshold (0.3): The fraction of frames with >5-pixel movement. Active touts average 0.45; passive waiting passengers average 0.12. The 0.3 threshold separates these populations with minimal overlap.
Combined score threshold (0.75): Tout classification requires a weighted combination of dwell, re-entry, engagement, and activity signals exceeding 0.75. The high threshold ensures passive long-dwell passengers (e.g., waiting for a seat) are not incorrectly filtered.

6.6 Ghost injection parameters

Association distance (120 pixels): Ghost injection matches lost tracks to visible heads within 120 pixels of the track's last position. This radius accommodates typical frame-to-frame movement plus localization error. Larger values caused incorrect associations in crowded scenes.
Injection confidence (0.6): Synthetic detections are assigned confidence 0.6, below the typical body detection confidence of 0.7–0.9. This ensures the tracker prefers real detections when available, using ghosts only to bridge gaps.
Maximum persistence (90 frames): Ghosts persist for at most 3 seconds. Longer persistence risks maintaining tracks for passengers who have actually exited the frame, while shorter values fail to bridge realistic occlusion durations observed in crowded boarding.
Body size estimation: Injected bodies are sized as 2× head width and 5× head height, approximating adult proportions. Adaptive sizing from head dimensions outperformed fixed body sizes, which produced poor IoU overlap when the tracker attempted to associate ghosts with reappearing real detections.

6.7 Motion validation parameters

Teleport ceiling (250 pixels): The maximum plausible single-frame displacement. At 30fps with typical camera geometry, this corresponds to approximately 2 meters per frame, exceeding any realistic pedestrian velocity. Movements beyond this threshold indicate tracking failures rather than genuine motion.
Rolling multiplier (6.0): Movements exceeding 6× the track's median historical step distance trigger teleport detection, even if below the absolute ceiling. This relative threshold catches ID switches where the new track location is implausible given the track's established movement pattern.
Crowd adjustment: The ceiling increases by 10 pixels per nearby person, accommodating the increased positional jitter caused by physical jostling in crowds. At density 10, the effective ceiling is 350 pixels.
Freeze duration (30 frames): Teleporting tracks are frozen for 1 second. If trajectory stabilizes (coefficient of variation <0.5 over 10 frames), the track is reinstated; otherwise, it is rejected. This window allows genuine tracks disrupted by momentary detection failures to recover while filtering persistent ID switches.

6.8 Door gating parameters

EMA smoothing window (5 frames): Door state predictions are smoothed over 5 frames to suppress detection noise. This introduces ~170ms latency, acceptable given that door transitions take 1–2 seconds.

Crossing grace period (15 frames): Tracks in the polygon when the door closes remain count-eligible for 0.5 seconds. This duration was measured from video as the typical time for a passenger to complete a crossing once committed to the threshold.

Hysteresis factor (0.7): State transitions require confidence exceeding 0.7× the opposing state's confidence, preventing rapid oscillation when the classifier is uncertain.

7 Evaluation

We evaluate DOORFUSE on the dataset described in Section 4, comparing against a standard tracking-based baseline. We report trip-level (per-video) counting results.

7.1 Experimental setup

7.1.1 Baseline. We compare against a line-crossing counter built on the same detection and tracking stack (YOLOv12 + BoT-SORT with ReID), but without DOORFUSE's behavioral verification. The baseline counts all track crossings of the counting line, applying only a minimum track lifetime filter (5 frames) to suppress spurious detections. This represents a strong baseline that already benefits from state-of-the-art tracking; improvements over this baseline isolate the contribution of doorway-specific mechanisms.

7.1.2 Evaluation protocol. Models are evaluated on the held-out test set (Section 4), with no parameter tuning on test clips. Trip-level counts (total boarding and alighting per video) are compared against third-party manual annotations. We report metrics separately for boarding (IN) and alighting (OUT) to expose directional biases.

7.1.3 Metrics. Let $\hat{c}_v$ and c_v denote predicted and ground-truth counts for video v, with V total videos.

Mean Absolute Error (MAE): $\frac{1}{V}\sum_v |\hat{c}_v - c_v|$. Average per-trip counting error in passengers.

Signed Error: $\sum_v(\hat{c}_v - c_v)$. Aggregate error across all trips; positive indicates overcounting, negative indicates undercounting.

Precision: Fraction of counted events corresponding to true passenger crossings, computed by matching predicted events to ground-truth annotations.

Recall: Fraction of true passenger crossings that were counted.

F1 Score: Harmonic mean of precision and recall.

7.1.4 Performance. Counting performance is reported in Table 5. DOORFUSE reduces entry MAE by 78% (5.56 → 1.20) and exit MAE by 85% (6.88 → 1.04). The baseline's systematic overcounting (+207 total) shifts to slight undercounting (−26), reflecting DOORFUSE's conservative design.

Kenya shows the largest improvements (entry MAE: 15.00 → 2.67), reflecting effective filtering of staff and tout crossings in single-door bidirectional traffic. Rwanda's separated doors present a simpler problem, yet MAE still drops 71–82%. Precision improves from 0.52–0.82 to 0.95–0.98 while recall remains high (0.90–0.98).

7.2 Ablation study

To isolate individual component contributions, we evaluate DOORFUSE variants on the clip subsets defined in Table 2. Each ablation disables one mechanism while keeping others fixed.

7.2.1 S1: Door-state gating. Ablation S1 results are reported in Table 6. Door-state gating restricts counting to intervals when the door is detected as open, suppressing false positives from closed-door queueing and threshold interactions.

Door-state gating reduces MAE by 44% for entries and 40% for exits, with corresponding F1 improvements of 5 and 12 percentage points respectively. The mechanism primarily improves precision by filtering spurious crossings during closed-door intervals. The increased undercounting for exits (signed error −6) reflects a precision-recall tradeoff: aggressive gating occasionally suppresses valid crossings that occur as the door closes, motivating the crossing grace period described in Section 5.3.

7.2.2 S2: Head-assisted tracking. Ablation S2 results are reported in Table 7. Head-assisted tracking addresses detection failures in crowded doorways by injecting ghost bodies when heads remain visible but full-body detections fail.

Head-assisted tracking yields the largest improvements of any single component. MAE drops by 60% for entries and 80% for exits; exit F1 improves by 34 % (0.50 → 0.84).

The baseline's poor exit performance reflects severe track fragmentation under crowding: when bodies occlude at the exit threshold, tracks break and re-initialize, causing the same passenger to be counted multiple times. The +42 signed error confirms systematic overcounting. Head-assisted tracking maintains identity through these occlusions, reducing exit overcounting by 83% (+42 → +7).

Entry counting shows a precision-recall tradeoff: the conservative ghost injection slightly undercounts (−12) compared to the baseline's near-neutral error (+2). This reflects the mechanism's design priority, preventing duplicates at the cost of occasionally missing crossings when head association fails.

7.2.3 S3: Staff and tout filtering. Ablation S3 results are reported in Table 8. Staff and tout filtering identifies persistent non-passengers through converging behavioral signals: side-dominant orientation, positional anchoring, pacing patterns, and polygon re-entry frequency.

Staff and tout filtering reduces aggregate overcounting by 74% (signed error +204 → +53). Entry overcounting drops dramatically from +99 to +7, a 93% reduction, indicating that staff and touts crossing the entrance threshold were the dominant source of false entry counts in these clips.

Exit counting also improves substantially, with signed error dropping 56% (+105 → +46). The smaller relative improvement for exits suggests additional overcounting sources beyond staff activity, such as track fragmentation during crowded alighting.

Entry F1 decreases slightly (0.72 → 0.69), reflecting a precision-recall tradeoff where the filtering mechanism occasionally misclassifies side-facing passengers as staff. This motivates the displacement override in our staff detector (Section 6): tracks with a displacement > 100 pixels are reclassified as passengers regardless of orientation.

Table 5: Counting performance by country, bus type, and direction. B = Baseline, D = DOORFUSE. Signed error shows systematic bias (positive = overcounting). Best results in bold.

			MAE↓		Precision↑		Recall↑		F1↑		Signed Err	
Country	**Mode**	**Category**	B	D	B	D	B	D	B	D	B	D
Kenya	B1+B2	Entries	15.00	**2.67**	0.52	**0.95**	0.88	**0.98**	0.68	**0.91**	+90	**-6**
Kenya	B1+B2	Exits	18.67	**2.50**	0.47	**0.98**	0.85	**0.90**	0.62	**0.91**	+102	**−13**
Rwanda	B1+B2	Entries	2.58	**0.74**	0.82	**0.97**	0.88	**0.94**	0.85	**0.95**	+11	**−6**
Rwanda	B1+B2	Exits	3.16	**0.58**	0.70	**0.95**	0.73	**0.94**	0.71	**0.95**	+4	**−1**
Both	Total	Entries	5.56	**1.2**	0.66	**0.96**	0.92	**0.92**	0.77	**0.94**	+101	**−12**
Both	Total	Exits	6.88	**1.04**	0.55	**0.96**	0.84	**0.90**	0.66	**0.93**	+106	**−14**

Table 6: Ablation S1: Effect of door-state gating. Base = gating disabled, Door = gating enabled.

	MAE↓		F1↑		Signed Err	
Category	Base	Door	Base	Door	Base	Door
Entries	4.86	**2.71**	0.88	**0.93**	−2	**−1**
Exits	6.71	**4.00**	0.66	**0.78**	**+1**	−6

Table 7: Ablation S2: Effect of head-assisted tracking. Base = body detection only, Dual = body + head with ghost injection.

	MAE↓		F1↑		Signed Err	
Category	Base	Dual	Base	Dual	Base	Dual
Entries	6.67	**2.67**	0.89	**0.95**	**+2**	−12
Exits	7.67	**1.50**	0.50	**0.84**	+42	**+7**

Table 8: Ablation S3: Effect of staff/tout filtering. Base = no filtering, Filter = behavioral filtering enabled.

	MAE↓		F1↑		Signed Err	
Category	Base	Filter	Base	Filter	Base	Filter
Entries	12.63	**10.63**	**0.72**	0.69	+99	**+7**
Exits	14.63	**10.75**	0.62	**0.65**	+105	**+46**

Table 9: Ablation S4: Effect of pose-based direction scoring. Base = trajectory-only, Pose = trajectory + orientation.

	MAE↓		F1↑		Signed Err	
Category	Base	Pose	Base	Pose	Base	Pose
Entries	12.71	**4.71**	0.70	**0.85**	+61	**−19**
Exits	16.57	**8.43**	0.60	**0.66**	+66	**−55**

7.2.4 S4: Pose-based direction scoring. Pose-based scoring uses shoulder orientation to determine crossing direction—boarding passengers typically face inward, alighting passengers face outward (Table 9). This signal is most valuable when traffic is bidirectional and trajectory alone is ambiguous.

Table 10: S5: Performance across visual modalities. Base = Baseline, Fuse = DOORFUSE.

		MAE↓		F1↑		Signed Err	
Modality	**Category**	Base	Fuse	Base	Fuse	Base	Fuse
Colored	Entries	1.20	**0.60**	0.95	**0.97**	+2	**−3**
	Exits	4.00	**1.00**	0.79	**0.96**	−20	**+1**
B&W / IR	Entries	4.50	**1.67**	0.85	**0.95**	−7	−8
	Exits	4.00	**0.67**	0.71	**0.94**	+8	**−4**
Modality Shift	Entries	21.25	**3.50**	0.67	**0.92**	+85	**−41**
	Exits	26.50	**2.50**	0.54	**0.92**	+106	**−8**

Entry MAE drops by 63% (12.71 → 4.71) and exit MAE by 49% (16.57 → 8.43). Entry F1 improves by 15 percentage points (0.70 → 0.85), the largest F1 gain across all ablations. The signed error shifts from overcounting (+61/+66) to undercounting (−19/−55), reflecting the scorer's conservative design: when orientation conflicts with trajectory evidence, the system defers rather than commits to a potentially incorrect count.

7.2.5 S5: Modality robustness (stress test). Bus CCTV systems frequently transition between color, B&W, and IR modes based on ambient lighting. We evaluate DOORFUSE across three conditions (Table 10): daytime color, nighttime B&W/IR, and mid-video modality transitions.

DOORFUSE maintains F1 0.94–0.97 across modalities. Modality shift clips present the greatest baseline challenge (signed error +191 from track fragmentation), yet DOORFUSE reduces MAE by 84–91% (entries: 21.25 → 3.50; exits: 26.50 → 2.50). The crowd-adaptive weighting compensates for reduced pose reliability by shifting weight to trajectory-based signals. Residual undercounting in modality shift clips (−41 entries) suggests future work on modality-invariant appearance features.

7.2.6 Summary. Table 11 ranks component contributions by their impact on the most relevant metric for each failure mode.

Head-assisted tracking provides the largest single improvement, addressing the track fragmentation that causes severe overcounting at crowded exits. Pose-based scoring is most valuable for bidirectional traffic where trajectory alone is ambiguous. Staff/tout

Table 11: Component contributions ranked by primary impact metric.

Component	Primary Metric	Baseline	Improvement
Head-assisted tracking	Exit F1	0.50	+0.34
Pose-based scoring	Entry F1	0.70	+0.15
Staff/tout filtering	Entry signed err	+99	−92
Door-state gating	Entry MAE	4.86	−44%

filtering targets a deployment-specific failure mode (persistent non-passengers) that dominates Kenya clips. Door-state gating provides consistent but smaller gains across all conditions.

The components address largely orthogonal failure modes: S2 targets occlusion-induced fragmentation, S3 targets non-passenger actors, S4 targets directional ambiguity, and S1 targets temporal false positives. This orthogonality explains why the full DOORFUSE system (combining all components) achieves larger gains than any single ablation.

7.3 Limitations and failure cases

Despite substantial improvements, DOORFUSE exhibits residual errors in several scenarios:

Modality transitions. Mid-video color-to-IR switches cause track fragmentation that behavioral verification cannot fully recover. The −41 entry signed error in S5 modality-shift clips reflects passengers lost during transitions. Modality-invariant appearance features may address this limitation.

Extreme crowding. When doorway density exceeds 12–15 persons, even head detection degrades due to severe overlap. The crowd-adaptive weighting helps but cannot compensate for complete detection failure.

Atypical passenger behavior. Passengers who exhibit staff-like behavior (e.g., standing side-facing while waiting for companions) are occasionally filtered. The displacement override catches most cases, but edge cases remain.

Door visibility. Clips where door state is not visible (e.g., cabin-facing cameras) cannot benefit from door-state gating. In these cases, DOORFUSE relies entirely on behavioral signals.

Stop-level attribution. While DOORFUSE accurately counts per-trip totals, attributing individual crossings to specific stops requires telematics alignment that introduces additional error sources not evaluated here.

8 Conclusion

We presented DOORFUSE, a door-aware multi-signal fusion pipeline that converts bus CCTV and telematics into stop-level boarding and alighting counts. By treating each crossing as a multi-evidence decision, combining trajectory, motion, orientation, and identity signals, DOORFUSE reduces the dominant failure modes of naive line-crossing: misses under occlusion, duplicates from oscillation, and spurious counts from staff and touts.

Across Nairobi and Kigali fleets with heterogeneous door layouts, viewpoints, and lighting regimes, DOORFUSE achieves 78–85% MAE reduction and F1 of 0.93–0.94. The resulting stop-indexed demand profiles enable planning analyses requiring finer granularity than trip totals: identifying high-demand stops, informal pickup locations, and service gaps. This approach offers low-resource transit agencies a practical path to high-quality passenger measurement using existing CCTV infrastructure.

References

[1] Nir Aharon, Roy Orfaig, and Danail Bobrovskis. 2022. BoT-SORT: Robust Associations for Multi-Pedestrian Tracking. *arXiv preprint arXiv:2206.14651* (2022). https://arxiv.org/abs/2206.14651

[2] Donatella Darsena, Giacinto Gelli, Ivan Iudice, and Francesco Verde. 2024. Sensing Technologies for Crowd Management, Adaptation, and Information Dissemination in Public Transportation Systems: A Review. *IEEE* (2024). Available in PDF format.

[3] Ana Belén Rodríguez González, Juan José Vinagre Díaz, and Mark Richard Wilby. 2020. Detailed origin-destination matrices of bus passengers using radio frequency identification. *IEEE Intelligent Transportation Systems Magazine* 14, 1 (2020), 141–152.

[4] Masood Jafari Kang, Shervin Ataeian, and SM Mahdi Amiripour. 2021. A procedure for public transit OD matrix generation using smart card transaction data. *Public Transport* 13, 1 (2021), 81–100.

[5] Chaoqun Kong, Tangyi Guo, and Liu He. 2021. Research on OD estimation of public transit passenger flow based on multi-source data. In *International Conference on Green Intelligent Transportation System and Safety*. Springer, 589–603.

[6] Anahid Nabavi Larijani, Ana-Maria Olteanu-Raimond, Julien Perret, Mathieu Brédif, and Cezary Ziemlicki. 2015. Investigating the mobile phone data to estimate the origin destination flow and analysis; case study: Paris region. *Transportation Research Procedia* 6 (2015), 64–78.

[7] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2024. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. arXiv:2303.05499 [cs.CV] https://arxiv.org/abs/2303.05499

[8] Xinyu Liu, Pascal Van Hentenryck, and Xilei Zhao. 2021. Optimization models for estimating transit network origin–destination flows with big transit data. *Journal of Big Data Analytics in Transportation* 3, 3 (2021), 247–262.

[9] Chris Mccarthy, Irene Moser, Prem Prakash Jayaraman, Hadi Ghaderi, Adin Ming Tan, and Ali Yavari. 2024. A Field Study of Internet of Things-Based Solutions for Automatic Passenger Counting. *IEEE Transactions on Intelligent Transportation Systems* (2024). Accessed via IEEE Xplore.

[10] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. 2022. Simple Open-Vocabulary Object Detection with Vision Transformers. arXiv:2205.06230 [cs.CV] https://arxiv.org/abs/2205.06230

[11] Kaan Ozbay, Neveen Shlayan, Hani Nassif, et al. 2017. Real-time estimation of transit OD patterns and delays using low cost-ubiquitous advanced technologies. (2017).

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:231591445

[13] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. 2018. CrowdHuman: A Benchmark for Detecting Human in a Crowd. *arXiv preprint arXiv:1805.00123* (2018).

[14] Andrew Zalewski, Daniel Sonenklar, Alexandra Cohen, Josie Kressner, and Gregory Macfarlane. 2019. *Public Transit Rider Origin-Destination Survey Methods and Technologies.* Technical Report. TCRP Synthesis, No. 138, Washington, DC. 170 pages. doi:10.17226/25428

[15] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. 2022. ByteTrack: Multi-Object Tracking by Associating Every Detection Box. (2022).

[16] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. 2019. Omni-Scale Feature Learning for Person Re-Identification. In *ICCV*. 3702–3712. https://openaccess.thecvf.com/content_ICCV_2019/html/Zhou_Omni-Scale_Feature_Learning_for_Person_Re-Identification_ICCV_2019_paper.html

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009