Bounding-Box Trajectories Matter for Video Anomaly Detection

Inpyo Song and Jangwon Lee

Sungkyunkwan University, Seoul, Republic of Korea
{songinpyo,leejang}@skku.edu

Abstract. Video anomaly detection is critical for public safety and security, yet remains highly challenging despite extensive research due to large variations in appearance, viewpoint, and scene dynamics. Among existing approaches, human pose-based methods have emerged as a major line of research, showing strong performance since many anomalies in public datasets involve humans and pose representations are robust to appearance changes while providing compact motion descriptions. However, these methods often overlook bounding-box trajectories, although such information is inherently available in pose-based pipelines. In this paper, we explicitly leverage these trajectories as a primary anomaly cue. We present TrajVAD, a framework that models multi-class bounding-box trajectories using normalizing flows to learn normal kinematic patterns. Its trajectory-only variant, TrajVAD-T, eliminates pose estimation, reaches 87.7 AP on ShanghaiTech, and achieves the best results on MSAD among compared methods. TrajVAD-P adds a reliability-gated pose branch and improves performance to 88.6 AUROC and 90.9 AP on ShanghaiTech, establishing bounding-box trajectories as an effective modality for video anomaly detection.

Keywords: Video anomaly detection · Bounding-box trajectories · Object-centric modeling · Normalizing flows

1 Introduction

Video anomaly detection (VAD) aims to discover events in videos that deviate from normal behavioral patterns. It serves as a key capability in intelligent surveillance and traffic monitoring systems, where a wide variety of dynamic scenes must be interpreted automatically [35–37]. Despite decades of progress, VAD remains challenging because scenes often contain diverse objects, including pedestrians, cyclists, and vehicles, and anomalous events can involve any of them [22, 48]. Moreover, practical deployment often demands real-time performance, as delayed detection offers limited utility in safety-critical environments. A practical VAD system must therefore generalize across multiple object classes while maintaining computational efficiency for real-time or large-scale operation.

Early deep learning approaches addressed this problem through pixel-level reconstruction or prediction [14, 22, 23, 27, 40, 43, 45]. While these methods capture global scene context, they inherently entangle task-relevant motion signals

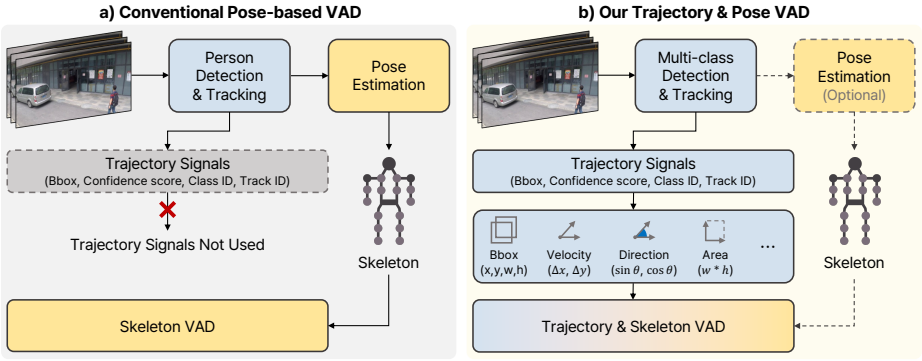

Fig. 1: Pose-based VAD methods (left) score anomalies from skeleton sequences and are limited to person-class tracks. TrajVAD (right) treats multi-class bounding-box trajectories as the primary signal, applicable to any detected object. TrajVAD-P adds an optional pose branch (dashed) activated only for human tracks when pose is reliable.

with extraneous appearance factors. Consequently, variations in lighting, background texture, or privacy-sensitive content often dominate the loss landscape, limiting the model's ability to focus purely on behavioral anomalies. To mitigate this sensitivity and decouple motion from appearance, pose-based methods have emerged as a popular alternative, modeling skeleton keypoint sequences to detect anomalies [5, 9, 16]. However, while effective in person-dominated scenes, these methods remain inherently human-centric, producing no anomaly cues for non-human entities such as vehicles and degrading significantly when keypoint estimation becomes unreliable under far-field views or heavy occlusion.

To address these limitations, we turn our attention to the bounding-box trajectory, an intermediate signal often overlooked in these frameworks. We observe that every pose-based pipeline begins with detection and tracking, producing per-instance bounding-box trajectories before skeleton extraction even starts. These trajectories thus naturally exist for every detected object class, not only persons, yet existing pipelines treat them merely as bookkeeping for identity association and continue to score anomalies solely from keypoints (Figure 1, left). Building on this observation, we revisit earlier attempts at trajectory-based anomaly detection. Although trajectory analysis has a history in VAD, prior works have either modeled only center-point paths [30] or coupled trajectories with skeleton predictions, thereby inheriting the person-specific constraints of pose methods [28, 29, 31, 38]. Consequently, the potential of multi-class bounding-box trajectories as a primary, standalone anomaly modality remains largely unexplored on modern VAD benchmarks.

To bridge this gap, we propose TrajVAD, a new framework that treats multi-class bounding-box trajectories as the primary anomaly signal and models their distribution through a likelihood-based objective trained solely on normal data (Figure 1, right). In the proposed framework, each tracked instance is represented as a sequence of trajectory-derived features, and a normalizing flow scores

anomalies via the negative log-likelihood under the learned normal distribution. TrajVAD features two variants that share the same trajectory backbone. TrajVAD-T operates purely on bounding-box trajectories and requires no pose estimation, while TrajVAD-P extends it with a reliability-gated pose branch activated only for human tracks when pose estimates are sufficiently confident.

We comprehensively evaluate the proposed TrajVAD framework on three widely used video anomaly detection benchmarks, ShanghaiTech, UBnormal, and MSAD, to validate its generality across diverse environments. On ShanghaiTech, TrajVAD-T remains competitive with pose-based baselines without pose estimation, showing that trajectory kinematics alone carry strong anomaly cues. TrajVAD-P further surpasses all compared pose-based methods on both AUROC and AP by incorporating the reliability-gated pose branch only for human tracks. Furthermore, leveraging multi-class trajectory coverage yields the largest gains over pose-based methods on MSAD, where non-human anomalies are prevalent.

Therefore, this work makes three key contributions:

- We establish multi-class bounding-box trajectories as a competitive primary modality for video anomaly detection, covering all detectable object classes without requiring pose estimation.
- We show that trajectory features, a free byproduct of the detection and tracking pipeline that pose-based methods already execute, can eliminate the pose estimation stage and its associated computational cost.
- We evaluate TrajVAD-T and TrajVAD-P on ShanghaiTech, UBnormal, and MSAD, where both variants run faster than pose-based methods and achieve the highest scores among all compared methods on ShanghaiTech and MSAD.

2 Related Work

2.1 Appearance-based Video Anomaly Detection

A major line of video anomaly detection models visual normality from RGB frames, optical flow, or deep visual features. Reconstruction-based methods learn to reproduce normal frames or latent visual representations with convolutional, recurrent, or memory-augmented autoencoders, and use reconstruction discrepancies as anomaly evidence [1,13,14,24]. Prediction-based methods learn temporal regularity by forecasting future frames, completing missing video events, or coupling optical-flow reconstruction with flow-guided frame prediction [22,23,44]. Recent approaches shift the scoring space from raw pixels to learned representations, using diffusion-based feature prediction, generative cooperative learning, or multiscale density estimation over visual features [27, 43, 45]. These frame-centric models preserve global scene context and can provide spatial evidence without requiring explicit detection or tracking. However, their scores are not naturally associated with persistent object instances, making object-wise temporal cues such as motion, scale variation, and changes in spatial extent less directly accessible.

2.2 Object-centric and Pose-based Video Anomaly Detection

A complementary paradigm, broadly termed object-centric approaches, grounds anomaly evidence in detected instances by attaching normality models to bounding box crops or tracklets [3, 7, 15, 33, 34, 41] or leveraging scene-graph relations [4,32]. These pipelines rely on detection and tracking to isolate objects, yet treat the resulting bounding-box trajectories merely as bookkeeping for identity association, overlooking the kinematic information they encode—translation, scale variation, and aspect-ratio dynamics.

Pose-based methods, a specialized subset, model each person as a skeleton keypoint sequence [5, 9, 10, 16, 17, 21, 25, 26, 28, 42], with architectures progressing from recurrent prediction [28] and graph clustering [26] through spatial-temporal graph convolutions [21,25] and normalizing flows [16] to diffusion models [9, 19]. While skeletons are compact and background-invariant, they are inherently human-centric, provide no signal for non-human objects, and become unreliable under occlusion or far-field views. Most critically, every pose-based pipeline already runs detection and tracking as a prerequisite, so per-instance bounding-box trajectories exist as a virtually **free** intermediate product, yet current methods overlook this signal in favor of the computationally expensive pose estimation stage.

2.3 Trajectory-based Video Anomaly Detection

Trajectory-level anomaly detection has a long history. Classical work models the distribution of center-point paths to flag routes that deviate from learned norms [30]. More recent methods extend center-point paths with multi-time scale trajectory prediction [31] and bi-directional trajectory prediction under pose constraints [18]. TrajREC [38] further introduces holistic multitask trajectory modeling by jointly reconstructing past, present, and future segments with a contrastive objective. TSGAD [29] includes a trajectory branch that predicts person-center coordinates obtained from the detector bounding box alongside a pose branch. However, in practice, these trajectory components remain closely tied to human-centric settings and person-level paths. Moreover, existing approaches like TrajREC and TSGAD do not extend their analysis to multi-class bounding-box kinematics, thereby neglecting non-person tracks and valuable dynamic signals such as scale evolution and detector confidence. Consequently, the potential of utilizing generic bounding-box trajectories from standard object trackers, which naturally cover all detected object categories, as a primary anomaly modality remains unexplored in video anomaly detection. TrajVAD addresses this gap by establishing trajectory-derived features as the core, class-agnostic input, while reserving pose estimation strictly as an optional refinement for human tracks.

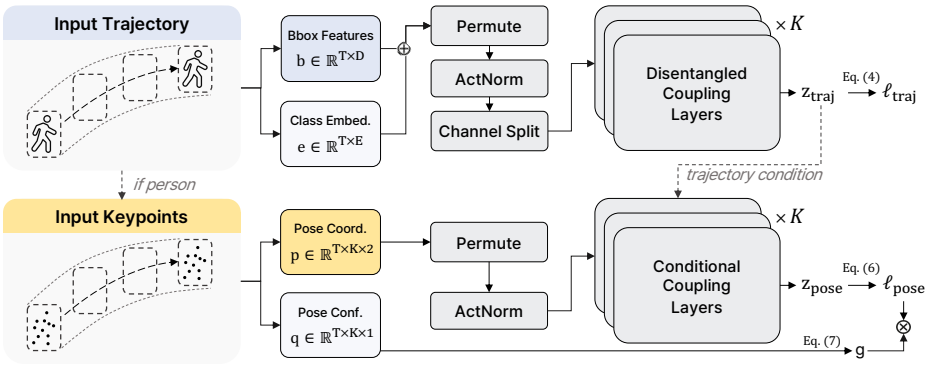

Fig. 2: TrajVAD pipeline. Multi-class tracks from detection and tracking are encoded as standardized trajectory-derived feature sequences and conditioned on class embeddings. TrajVAD-T (top row) maps segments through a normalizing flow and uses the negative log-likelihood as the anomaly score. TrajVAD-P (both rows) adds a pose branch conditioned on the trajectory latent $\mathbf{z}_{\text{traj}}$, gated by pose reliability g.

3 Method

3.1 Overview

We present a framework that captures normal motion patterns directly from multi-class bounding-box trajectories. Given an input video, the pipeline first employs an object detector and a multi-class tracker to generate per-instance tracks. These tracks are encoded as sequences of trajectory-derived features, augmented with learnable class embeddings to capture category-specific dynamics. To quantify normality, a normalizing flow model learns the distribution of these features and assigns a likelihood score to each track segment.

We propose two variants: TrajVAD-T, which relies solely on trajectory data, and TrajVAD-P, which extends the former with a pose branch to refine human tracks when skeleton estimates are available. Both variants share a unified trajectory backbone and scoring mechanism, as illustrated in Figure 2.

3.2 Trajectory Representation

Feature extraction. To model motion dynamics effectively, we begin with the bounding-box sequences produced by the tracker. Each box encodes both spatial position and size at every frame but, in its raw form, does not yet capture temporal dependencies. We normalize all box coordinates, including center position (c_x, c_y) and dimensions (w, h), to $[0, 1]$ by dividing by frame resolution, making features resolution-independent. From these normalized quantities, we construct a feature vector $\mathbf{b}_t \in \mathbb{R}^D$ at each frame. The features are organized into six groups (Table 1). State features record instantaneous position and box geometry. Temporal dynamics capture velocity, acceleration, and higher-order

Table 1: Bounding-box trajectory features (27 total). Coordinates and dimensions are normalized to $[0, 1]$ by frame resolution before feature computation.

Group	Features
State (6)	Center (c_x, c_y), box size (w, h, area, ratio)
Temporal dynamics (10)	Velocity (v_x, v_y, speed), acceleration (a_x, a_y, acc), direction ($\sin\theta$, $\cos\theta$), jerk, curvature
Geometric dynamics (2)	Box expansion rate, aspect-ratio velocity
Pseudo-physical (2)	Kinetic energy proxy (area $\times$ speed^2), path efficiency
Perspective-normalized (6)	Speed per box dim (v/h, v/w, v/area), acceleration per box dim (a/h, a/w, a/area)
Confidence (1)	Detector confidence (conf)

derivatives of center motion. Geometric dynamics track how box shape changes over time. Pseudo-physical features combine size and speed into energy and efficiency proxies. Perspective-normalized features divide translational quantities by box dimensions to reduce dependence on camera-to-subject distance. Finally, detector confidence provides a scalar signal that correlates with occlusion and unusual appearance. All features are standardized using training-split statistics. Exact definitions for each feature are provided in the supplementary material. To ensure stable measurements of motion, the bounding-box coordinates from the tracker are smoothed before feature extraction. Each track is then segmented into overlapping windows of T frames, and segments shorter than T are discarded.

Class embedding. Each trajectory segment is further contextualized by its associated object class. Because the trajectory features defined above rely exclusively on generic bounding-box statistics, they naturally generalize across all detected object categories, including but not limited to the pedestrian and person classes. We associate each segment with a class label $c \in \{0, \ldots, C-1\}$. We then retrieve a trainable embedding $\mathbf{e} \in \mathbb{R}^E$ for c and concatenate it to the feature vector at every time step, yielding

$$\mathbf{x}_t = [\mathbf{b}_t \parallel \mathbf{e}] \in \mathbb{R}^{D+E}. \tag{1}$$

This allows a single shared flow to distinguish class-specific normality patterns.

3.3 TrajVAD-T: Trajectory Normalizing Flow

TrajVAD-T models the distribution of normal trajectory segments with a normalizing flow. The flow maps a segment $\mathbf{x} \in \mathbb{R}^{T\times(D+E)}$ to a latent vector $\mathbf{z}_{\text{traj}}$

through a sequence of invertible transforms and evaluates log-likelihood under a Gaussian prior.

The first transform is ActNorm [20], a data-adaptive affine layer initialized from the first training batch:

$$\mathbf{z}_0 = (\mathbf{x} + \boldsymbol{\beta}) \odot \boldsymbol{\sigma} \tag{2}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$ are learned channel-wise bias and scale parameters. The normalized representation $\mathbf{z}_0$ then passes through K alternating affine coupling layers [6]. Each layer splits the feature channels into two halves $(\mathbf{u}, \mathbf{v})$ and transforms one half while keeping the other fixed:

$$\mathbf{u}' = \mathbf{u}, \qquad \mathbf{v}' = (\mathbf{v} + t(\mathbf{u})) \odot \mathbf{a}(\mathbf{u}), \quad \mathbf{a}(\mathbf{u}) = \mathrm{sigmoid}(s(\mathbf{u}) + 2) + \epsilon \tag{3}$$

where a causal convolutional subnetwork predicts the translation $t(\cdot)$ and raw scale $s(\cdot)$ parameters. The subnet uses causal 1-D convolutions over the temporal axis with increasing dilation to expand the receptive field within the segment. Alternating layers swap which half is transformed, ensuring all channels interact across the stack. The final output latent is $\mathbf{z}_{\mathrm{traj}} = f_K \circ \cdots \circ f_1(\mathbf{z}_0)$.

The unnormalized log-likelihood of a trajectory segment is

$$\ell_{\mathrm{traj}} = \log p(\mathbf{z}_{\mathrm{traj}}) + \log\left|\det J\right| \tag{4}$$

where $p(\mathbf{z}_{\mathrm{traj}}) = \mathcal{N}(\mathbf{z}_{\mathrm{traj}}; \mu_0\mathbf{1}, \mathbf{I})$ with a fixed offset μ_0, and $\log|\det J|$ accumulates log-determinants from ActNorm and all coupling layers. The training objective minimizes the segment-normalized negative log-likelihood:

$$\mathcal{L}_{\mathrm{traj}} = -\frac{\ell_{\mathrm{traj}}}{T(D+E)}. \tag{5}$$

At test time, $\mathcal{L}_{\mathrm{traj}}$ serves as the per-segment anomaly score.

3.4 TrajVAD-P: Reliability-Gated Pose Branch

TrajVAD-P extends TrajVAD-T with an optional pose branch for person-class segments that have a valid pose estimate. We extract 17-keypoint poses using AlphaPose [8] and split each estimate into keypoint coordinates $\mathbf{p} \in \mathbb{R}^{T\times 2J}$ and per-keypoint confidence scores $\mathbf{q} \in \mathbb{R}^{T\times J}$ (J=17), both aligned to the same track window.

The pose branch is a separate normalizing flow with its own ActNorm and K_p coupling layers. Each pose coupling layer conditions its scale and translation subnets on the trajectory latent $\mathbf{z}_{\mathrm{traj}}$. The pose flow therefore models pose normality conditioned on the trajectory cues. The pose log-likelihood is computed analogously to the trajectory branch:

$$\ell_{\mathrm{pose}} = \log p(\mathbf{z}_{\mathrm{pose}}) + \log\left|\det J_{\mathrm{pose}}\right|. \tag{6}$$

The pose branch activates only when the segment belongs to the person class (g_{cls}=1) and a valid pose estimate exists (g_{valid}=1). To prevent unreliable pose

from corrupting the trajectory score, the pose contribution is further scaled by the mean keypoint confidence $\bar{q} \in [0, 1]$. These three conditions are combined into a single scalar gate:

$$g = g_{\text{cls}} \cdot g_{\text{valid}} \cdot \bar{q}. \tag{7}$$

When $g = 0$, TrajVAD-P reduces exactly to TrajVAD-T.

The combined score normalizes the joint log-likelihood by the effective dimensionality:

$$\mathcal{L}_{\text{total}} = -\frac{\ell_{\text{traj}} + \lambda\, g\, \ell_{\text{pose}}}{T \cdot (D_{\text{traj}} + D_{\text{pose}} \cdot g)} \tag{8}$$

where $\ell_{\text{traj}} = \log p(\mathbf{z}_{\text{traj}}) + \log|\det J_{\text{traj}}|$ and ℓ_{pose} are the unnormalized log-likelihoods of the trajectory and pose flows respectively, λ is a pose weight hyperparameter, and D_{traj}, D_{pose} are the trajectory and pose feature dimensions. The denominator $T \cdot (D_{\text{traj}} + D_{\text{pose}} \cdot g)$ normalizes by the effective dimensionality, so that the score scale remains consistent whether pose is active or not. When pose is inactive ($g = 0$), this reduces to $\mathcal{L}_{\text{traj}}$ (Eq. 5).

4 Experimental Results

4.1 Datasets

We evaluate the proposed approach on three public VAD benchmarks spanning real and synthetic surveillance scenarios.

ShanghaiTech [22] contains 330 normal training videos and 107 test videos recorded by 13 fixed cameras on a university campus. Abnormal events include running, fighting, skateboarding, and cycling, and 6 test videos feature non-human anomalies (e.g., vehicles entering pedestrian areas).

UBnormal [2] is a synthetic benchmark rendered with Cinema4D across 29 scenes, containing 268 training, 64 validation, and 211 test videos with both frame- and pixel-level annotations. It includes normal actions such as walking, talking, and sitting, and staged anomalies such as falling, lying down, stealing, and dancing.

MSAD [48] is the most diverse benchmark considered here, with 720 videos covering 55 abnormal event types across varied locations and viewpoints. It contains many non-human anomalies (e.g., industrial and traffic events) that yield no skeleton keypoints. We additionally consider its Human-Related subset (MSAD-HR), which includes seven categories (e.g., Fighting, Robbery, and Traffic Accident) and often involves non-person objects alongside humans, making it our main benchmark for multi-class trajectory coverage.

HR subsets. Since skeleton-based methods provide no anomaly signal for non-human events, we follow prior protocols [5, 16, 28, 48] and define Human-Related (HR) subsets for all three datasets.

4.2 Evaluation Metrics

We assign an anomaly score to each frame in the test set, concatenate scores across all test videos, and compute the Area Under the Receiver Operating Characteristic curve (AUROC) and Average Precision (AP) against the binary ground-truth labels. This micro-averaged protocol follows prior work [5, 16].

4.3 Implementation Details

We use YOLOX [11] for detection and OSNet [47] for re-identification, following STG-NF [16] and SeeKer [5]. Unlike these pose-based baselines, we replace the pose-specific tracker with ByteTrack [46] to support multi-class tracking beyond persons. Tracks are segmented into windows of T=12, 16, 24 for UBnormal, ShanghaiTech, and MSAD, respectively. Before windowing, bounding-box coordinates are smoothed with a Kalman filter. Missing detections spanning at most 10 frames are filled by linear interpolation, whereas longer gaps split the track. This preprocessing handles transient detector dropouts without forcing long occluded intervals into a single trajectory. The trajectory flow uses K=6 coupling layers for TrajVAD-T and K=18 for TrajVAD-P, with a 3-dimensional class embedding over the 80 COCO classes. The Gaussian prior uses a mean offset μ_0=3. Further implementation details are provided in the supplementary material.

4.4 Comparison with State-of-the-Art Methods

ShanghaiTech [22]. Table 2 reports results on ShanghaiTech. ShanghaiTech anomalies are predominantly motion-defined events such as running, fighting, and skateboarding in pedestrian areas, where translational dynamics and scale changes in bounding boxes provide direct discriminative signal. TrajVAD-T trails STG-NF by 1.0 AUROC point but achieves 87.7% AP, exceeding STG-NF and SeeKer by 10.1 and 7.7 AP points, respectively. TrajVAD-P achieves the highest AUROC of 88.6 on both ShanghaiTech and ShanghaiTech-HR, with AP of 90.9% and 91.1% respectively, surpassing all compared methods on both metrics. This gain indicates that pose cues provide complementary information for human-centric anomalies whose body configuration changes are not fully captured by bounding-box kinematics.

UBnormal [2]. UBnormal presents a harder setting for trajectory-based detection (Table 3). TrajVAD-P reaches 73.8 AUROC and 68.3 AP on UBnormal, surpassing STG-NF and all other compared pose-based methods except SeeKer, but trailing SeeKer by 4.1 AUROC points. TrajVAD-T achieves competitive AUROC of 68.0 without pose estimation.

This gap reflects a challenge specific to UBnormal's synthetic domain. We observe that rendered sequences produce unstable object detection, with class labels flipping between frames and confidence scores that are noisy and low. The feature ablation in Table 6 shows that detector confidence is the most impactful

Table 2: Comparison with RGB, flow, multimodal, pose-based, and trajectory-based methods on ShanghaiTech [22]. SHT-HR denotes the human-related subset. The supplementary material reports the property breakdown and localization metrics.

Method	Venue	AUROC		AP	
		SHT	SHT-HR	SHT	SHT-HR
RGB, flow, and multimodal baselines					
sRNN [24]	ICCV'17	68.0	–	–	–
Conv-AE [14]	CVPR'16	70.4	69.8	–	–
LSA [1]	CVPR'19	72.5	–	–	–
FFP [22]	CVPR'18	72.8	72.7	–	–
VEC [44]	ACM MM'20	74.8	–	–	–
HF^2 [23]	ICCV'21	76.2	–	–	–
FPDM [43]	ICCV'23	78.6	–	–	–
CAE-SVM [39]	ICCV'17	78.7	–	–	–
GCL [45]	CVPR'22	79.6	–	–	–
BA-AED [12]	TPAMI'22	82.7	–	–	–
SSMTL++ [3]	CVIU'23	83.8	–	–	–
Jigsaw [40]	ECCV'22	84.2	84.7	–	–
MULDE [27]	CVPR'24	<u>86.7</u>	–	–	–
Pose and trajectory baselines					
BiPOCO [18]	arXiv'22	–	74.9	–	–
MPED-RNN [28]	CVPR'19	73.4	75.4	–	–
MTP [31]	WACV'20	76.0	77.0	–	–
GEPC [26]	CVPR'20	76.1	74.8	–	–
PoseCVAE [17]	ICPR'21	–	75.5	–	–
Normal Graph [25]	NC'21	–	76.5	–	–
COSKAD [10]	PR'24	–	77.1	–	–
GCAE-LSTM [21]	NC'22	–	77.2	–	–
MoCoDAD [9]	ICCV'23	–	77.6	–	–
TrajREC [38]	WACV'24	–	77.9	–	–
TSGAD-Traj [29]	WACV'24	67.8	68.4	61.2	61.3
TSGAD [29]	WACV'24	80.6	81.5	73.9	74.2
GiCiSAD [19]	WACV'25	–	78.0	–	–
STG-NF [16]	ICCV'23	85.9	<u>87.4</u>	77.6	81.4
SeeKer [5]	ICCV'25	85.5	86.9	80.0	81.5
TrajVAD-T	–	84.9	84.1	<u>87.7</u>	<u>87.5</u>
TrajVAD-P	–	**88.6**	**88.6**	**90.9**	**91.1**

single feature for TrajVAD-T (−2.4 AUROC when removed), so noisy confidence on synthetic imagery directly undermines the trajectory representation. Adding pose in TrajVAD-P recovers 5.8 AUROC points over TrajVAD-T, yet cannot fully compensate for the weakened trajectory base. A failure case analysis illustrating these detection instabilities is provided in the supplementary material.

MSAD [48]. MSAD tests whether multi-class trajectory coverage translates to measurable gains over pose-only methods (Table 4). Unlike ShanghaiTech and UBnormal, where anomalies are predominantly motion deviations such as running or jumping, MSAD includes semantically complex categories such as robbery and vandalism, and MSAD-HR contains classes like traffic accident where the primary object is a vehicle.

On MSAD-HR, TrajVAD-T achieves 69.7 AUROC and 60.4 AP, exceeding SeeKer by 8.6 AUROC points and STG-NF by 14.0 AUROC points. This is the largest AUROC gain among the three benchmarks. Because pose-based

Table 3: Comparison with RGB, flow, multimodal, pose-based, and trajectory-based methods on UBnormal [2]. UBn-HR denotes the human-related subset. The supplementary material reports the property breakdown and localization metrics.

Method	Venue	AUROC		AP	
		UBn	UBn-HR	UBn	UBn-HR
RGB, flow, and multimodal baselines					
Jigsaw [40]	ECCV'22	56.4	–	–	–
BA-AED [12]	TPAMI'22	59.3	–	–	–
SSMTL++ [3]	CVIU'23	62.1	–	–	–
FPDM [43]	ICCV'23	62.7	–	–	–
MULDE [27]	CVPR'24	72.8	–	–	–
Pose and trajectory baselines					
BiPOCO [18]	arXiv'22	50.7	52.3	–	–
GEPC [26]	CVPR'20	53.4	55.2	–	–
MPED-RNN [28]	CVPR'19	60.6	61.2	–	–
COSKAD [10]	PR'24	65.0	65.5	–	–
TrajREC [38]	WACV'24	68.0	68.2	–	–
GiCiSAD [19]	WACV'25	68.6	68.8	–	–
MoCoDAD [9]	ICCV'23	68.3	68.4	–	–
STG-NF [16]	ICCV'23	71.8	71.5	62.7	67.2
SeeKer [5]	ICCV'25	**77.9**	**78.9**	**80.3**	**79.8**
TrajVAD-T	–	68.0	68.0	63.2	63.4
TrajVAD-P	–	<u>73.8</u>	<u>73.8</u>	<u>68.3</u>	<u>68.4</u>

models produce no signal for non-human objects, TrajVAD's uniform modeling of bounding-box kinematics across all tracked classes provides a structural advantage when non-person anomalies are prevalent. TrajVAD-T outperforms TrajVAD-P on both MSAD-HR and MSAD, consistent with the expectation that human-pose gating provides limited benefit when the dominant anomalous objects are non-human.

On full MSAD, performance drops for all methods because categories such as Explosion and Fire produce no trackable objects; TrajVAD-T still leads with 57.2 AUROC ahead of STG-NF (53.8), but the 12.5-point gap from MSAD-HR reflects anomaly types beyond the detector's reach.

Overall, the three benchmarks highlight complementary roles of trajectory and pose cues. Trajectory-only modeling is effective when anomalies are expressed through object-level motion or involve non-human objects, as in MSAD. Pose augmentation is most useful for human-centric anomalies with informative body-configuration changes, as in ShanghaiTech. UBnormal, in contrast, exposes the sensitivity of trajectory modeling to synthetic detector noise.

4.5 Computational Cost

Table 5 breaks down the end-to-end cost per temporal window (segment). Pose-based methods use a person-specific tracker, while TrajVAD uses a general bounding-box tracker (ByteTrack [46]). Both trackers run in roughly 1 ms per frame, so the detection and re-identification stages (YOLOX [11] + OSNet [47]) dominate preprocessing cost. This explains the negligible gap between the two pipelines (105.8 vs. 104.4 ms/frame). TrajVAD-T skips pose estimation entirely,

Table 4: Comparison with pose-based and trajectory-based methods on MSAD [48]. MSAD-HR denotes the human-related subset.

Method	Venue	AUROC		AP	
		MSAD	MSAD-HR	MSAD	MSAD-HR
MoCoDAD [9]	ICCV'23	–	53.9	–	51.7
STG-NF [16]	ICCV'23	53.8	55.7	37.5	56.5
TrajREC [38]	WACV'24	51.2	54.8	39.8	53.2
SeeKer [5]	ICCV'25	–	61.1	–	<u>60.1</u>
TrajVAD-T	–	**57.2**	**69.7**	**42.7**	**60.4**
TrajVAD-P	–	<u>55.5</u>	<u>68.5</u>	<u>41.5</u>	58.5

Table 5: Computational cost breakdown. Preprocessing is per frame, and inference is per segment. Total includes preprocessing and inference because stride-1 sliding-window evaluation produces one new segment per frame.

Method	Preprocess (ms/frame)		Inference	Total	SHT-HR	
	Det+Track	+Pose	(ms/seg)	(ms/seg)	AUROC	AP
TrajREC [38]	105.8	31.0	4.5	<u>141.3</u>	77.9	–
STG-NF [16]	105.8	31.0	8.8	145.6	<u>87.4</u>	81.4
MoCoDAD [9]	105.8	31.0	2,857.0	2,993.8	77.6	–
GiCiSAD [19]	105.8	31.0	1,296.1	1,432.9	78.0	–
TrajVAD-T	104.4	–	2.8	**107.2**	84.1	<u>87.5</u>
TrajVAD-P	104.4	31.0	8.6	<u>144.0</u>	**88.6**	**91.1**

saving 31 ms per frame compared to methods that require skeleton extraction. MoCoDAD and GiCiSAD use diffusion-based inference, which accounts for over 90% of their total cost. The hardware configuration used for speed benchmarking is provided in the supplementary material.

4.6 Ablation Study

Feature group ablation. Table 6 evaluates each trajectory feature group through a leave-one-group-out protocol on ShanghaiTech. Every group contributes positively for TrajVAD-T, confirming that the full 27-dimensional representation is well-utilized.

Detector confidence is the most impactful single feature for both variants (-2.4 ShanghaiTech AUROC for TrajVAD-T, -2.7 for TrajVAD-P). Unlike all other features derived from bounding-box coordinates, confidence is the only signal sourced from the detector's internal state, encoding detection quality under occlusion, truncation, and unusual appearance that no kinematic feature can recover, which explains why removing a single scalar causes the largest drop. For TrajVAD-T, state features show the smallest effect (-0.3) because higher-order derivatives already subsume raw position and geometry. Perspective-normalized

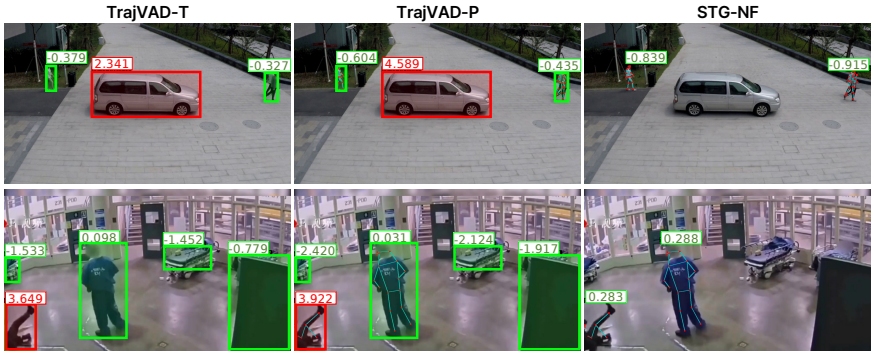

Fig. 3: Qualitative comparison on ShanghaiTech and MSAD. Red boxes and anomaly scores (higher means anomaly) indicate detected anomalies. Top: a car in a pedestrian zone is invisible to pose-based STG-NF but flagged by TrajVAD through bounding-box kinematics. Bottom: partial occlusion corrupts skeleton estimation, suppressing the STG-NF anomaly score, while TrajVAD maintains detection from trajectory features.

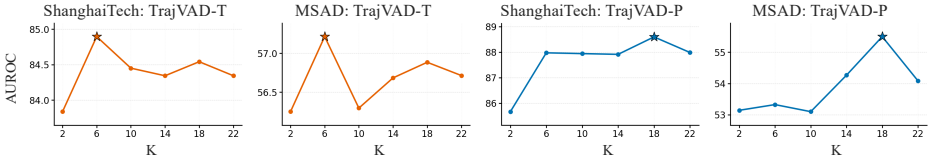

Fig. 4: Effect of flow depth K on AUROC for TrajVAD-T and TrajVAD-P on ShanghaiTech and MSAD. Stars mark the best K per panel. TrajVAD-T is robust across depths, while TrajVAD-P peaks at K=18.

features are the second most impactful group for TrajVAD-T (-1.5) but cause no degradation for TrajVAD-P ($+0.0$), as the pose branch's pelvis normalization already encodes body-relative scale, making perspective correction redundant when pose is available.

Flow depth ablation. Figure 4 varies the number of coupling layers K from 2 to 22 on ShanghaiTech and MSAD. TrajVAD-T is relatively insensitive to K on ShanghaiTech, with AUROC ranging from 83.8 to 84.9 across the sweep and the best result at K=6. On MSAD, TrajVAD-T shows a similar pattern, peaking at K=6 with 57.2 AUROC. TrajVAD-P exhibits a clearer optimum at K=18 on both datasets, suggesting that the pose-conditioned flow benefits from additional depth to model the joint trajectory-pose distribution. Based on these results, TrajVAD-T uses K=6 and TrajVAD-P uses K=18 in the main comparison tables.

Robustness to detection failures. To evaluate robustness to upstream detection errors, we randomly remove 10% of detections per track on ShanghaiTech

Table 6: Feature group ablation on ShanghaiTech (leave-one-group-out, AUROC %). Each row removes one feature group from the full 27-dimensional set. Feature groups follow Table 1. D is the remaining feature dimensionality. Δ denotes AUROC change relative to the full set.

Features	D	TrajVAD-T		TrajVAD-P	
		SHT (Δ)	SHT-HR (Δ)	SHT (Δ)	SHT-HR (Δ)
Full 27D	27	84.9	84.1	88.6	88.6
− State (6)	21	84.6 **−0.3**	83.8 **−0.3**	87.5 **−1.1**	87.4 **−1.2**
− Temporal (10)	17	83.9 **−1.0**	83.0 **−1.1**	88.2 **−0.4**	88.0 **−0.6**
− Geometric (2)	25	84.0 **−0.9**	83.1 **−1.0**	88.0 **−0.6**	87.8 **−0.8**
− Pseudo-physical (2)	25	84.1 **−0.8**	83.2 **−0.9**	88.0 **−0.6**	87.7 **−0.9**
− Perspective (6)	21	83.4 **−1.5**	82.6 **−1.5**	88.6 **+0.0**	88.6 **+0.0**
− Confidence (1)	26	82.5 **−2.4**	81.9 **−2.2**	85.9 **−2.7**	85.8 **−2.8**

Table 7: Robustness to simulated missed detections on ShanghaiTech. AUROC (%) is reported under clean detections and after randomly removing 10% of detections per track. Δ denotes the change from the clean setting.

Method	Clean	10% removed	Δ
TSGAD [29]	80.6	76.1	−4.5
STG-NF [16]	85.9	80.7	−5.2
TrajVAD-T	84.9	82.4	**−2.5**

and apply the preprocessing in Sec. 4.3. Table 7 shows that TrajVAD-T loses 2.5 AUROC points, while STG-NF and TSGAD lose 5.2 and 4.5 points, respectively. These results indicate that bounding-box trajectory modeling remains stable under short detection gaps, especially when combined with Kalman smoothing, interpolation, and conservative track splitting.

4.7 Qualitative Analysis

Figure 3 shows two representative failure modes of pose-based detection that TrajVAD handles correctly. In the top row, a car drives through a pedestrian area. STG-NF assigns no anomaly score to the vehicle because pose estimation is undefined for non-person objects, whereas both TrajVAD-T and TrajVAD-P flag the car through its bounding-box trajectory alone. In the bottom row, a person is only partially visible with legs occluded, causing skeleton estimation to fail. STG-NF produces a low anomaly score from the corrupted pose, while TrajVAD-T still detects the anomaly from bounding-box kinematics and TrajVAD-P further amplifies the score by gating out the unreliable pose.

5 Limitations and Future Work

One limitation of the proposed method is that TrajVAD's performance is strongly affected by the quality of its upstream detector. As shown on UBnormal (Table 3), noisy confidence scores and class-label flipping in synthetic imagery degrade the trajectory representation. The feature ablation in Table 6 also shows that detector confidence is the most influential feature, with a -2.4 AUROC drop when it is removed. Thus, scenes with poorly calibrated confidence scores, missed detections, or unstable track IDs can reduce TrajVAD's accuracy. Another limitation stems from the detector's class coverage. Although the shared COCO-trained detector spans 80 categories, which is substantially broader than person-only pose pipelines, anomalies involving out-of-vocabulary objects or events that yield no foreground detection remain beyond reach. The 12.5-point gap between MSAD-HR (69.7 AUROC) and MSAD (57.2 AUROC) quantifies this limitation. Adopting an open-world or class-agnostic detector would extend this coverage without altering the trajectory modeling itself.

6 Conclusion

This paper revisited bounding-box trajectories as a primary modality for video anomaly detection. Although these trajectories are already produced by detection and tracking pipelines, they are typically used only for identity association and rarely treated as anomaly evidence. We showed that multi-class bounding-box dynamics provide strong object-level cues, enabling competitive anomaly detection without dense appearance modeling or mandatory pose estimation. TrajVAD represents each tracked object with class-aware trajectory features and models normality with a normalizing flow. The trajectory-only variant removes the pose estimation stage while retaining strong detection accuracy and improving efficiency over pose-dependent pipelines. The reliability-gated pose variant further complements trajectory cues in human-centric cases, where body configuration provides information beyond bounding-box motion. Experimental results on three widely used VAD benchmarks validated the proposed design. In particular, multi-class trajectory coverage yielded the largest gains on MSAD, whereas reliability-gated pose augmentation achieved the best performance on ShanghaiTech, demonstrating the complementary benefits of the two variants. These findings establish bounding-box trajectories as a simple, scalable, and previously underutilized modality for VAD.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) (No. RS-2024-00458696), the MSIT (Ministry of Science and ICT), Korea, under the Global Scholars Invitation Program (RS-2024-00459638), and the Graduate School of Virtual Convergence support program (IITP-2026-RS-2023-00254129), both supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

References

1. Abati, D., Porrello, A., Calderara, S., Cucchiara, R.: Latent space autoregression for novelty detection. In: CVPR. pp. 481–490 (2019)
2. Acsintoae, A., Florescu, A., Georgescu, M.I., Mare, T., Sumedrea, P., Ionescu, R.T., Khan, F.S., Shah, M.: UBnormal: New benchmark for supervised open-set video anomaly detection. In: CVPR. pp. 20143–20153 (2022)
3. Barbalau, A., Ionescu, R.T., Georgescu, M.I., Dueholm, J., Ramachandra, B., Nasrollahi, K., Khan, F.S., Moeslund, T.B., Shah, M.: SSMTL++: Revisiting self-supervised multi-task learning for video anomaly detection. Computer Vision and Image Understanding **229**, 103656 (2023)
4. Dawoud, K., Zaheer, Z., Khan, M., Nandakumar, K., Elsaddik, A., Khan, M.H.: FusedVision: A knowledge-infusing approach for practical anomaly detection in real-world surveillance videos. In: CVPR. pp. 4036–4046 (2025)
5. Delić, A., Grcic, M., Šegvić, S.: Sequential keypoint density estimator: an overlooked baseline of skeleton-based video anomaly detection. In: ICCV. pp. 11579–11589 (2025)
6. Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using Real NVP. arXiv preprint arXiv:1605.08803 (2016)
7. Doshi, K., Yilmaz, Y.: Towards interpretable video anomaly detection. In: WACV. pp. 2655–2664 (2023)
8. Fang, H.S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.L., Lu, C.: AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time. IEEE TPAMI **45**(6), 7157–7173 (2022)
9. Flaborea, A., Collorone, L., Di Melendugno, G.M.D., D'Arrigo, S., Prenkaj, B., Galasso, F.: Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection. In: ICCV. pp. 10318–10329 (2023)
10. Flaborea, A., di Melendugno, G.M.D., D'Arrigo, S., Sterpa, M.A., Sampieri, A., Galasso, F.: Contracting skeletal kinematics for human-related video anomaly detection. PR **156**, 110817 (2024)
11. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: YOLOX: Exceeding YOLO series in 2021. arXiv preprint arXiv:2107.08430 (2021)
12. Georgescu, M.I., Ionescu, R.T., Khan, F.S., Popescu, M., Shah, M.: A background-agnostic framework with adversarial training for abnormal event detection in video. IEEE TPAMI **44**(9), 4505–4523 (2021)
13. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Hengel, A.v.d.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: ICCV. pp. 1705–1714 (2019)
14. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S.: Learning temporal regularity in video sequences. In: CVPR. pp. 733–742 (2016)
15. Hinami, R., Mei, T., Satoh, S.: Joint detection and recounting of abnormal events by learning deep generic knowledge. In: ICCV. pp. 3619–3627 (2017)
16. Hirschorn, O., Avidan, S.: Normalizing flows for human pose anomaly detection. In: ICCV. pp. 13545–13554 (2023)
17. Jain, Y., Sharma, A.K., Velmurugan, R., Banerjee, B.: Posecvae: Anomalous human activity detection. In: ICPR. pp. 2927–2934 (2021)
18. Kanu-Asiegbu, A.M., Vasudevan, R., Du, X.: BiPOCO: Bi-directional trajectory prediction with pose constraints for pedestrian anomaly detection. arXiv preprint arXiv:2207.02281 (2022)

19. Karami, A., Ho, T.K.K., Armanfard, N.: Graph-jigsaw conditioned diffusion model for skeleton-based video anomaly detection. In: WACV. pp. 4237–4247 (2025)
20. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. NeurIPS **31** (2018)
21. Li, N., Chang, F., Liu, C.: Human-related anomalous event detection via spatial-temporal graph convolutional autoencoder with embedded long short-term memory network. Neurocomputing **490**, 482–494 (2022)
22. Liu, W., Luo, W., Lian, D., Gao, S.: Future frame prediction for anomaly detection–a new baseline. In: CVPR. pp. 6536–6545 (2018)
23. Liu, Z., Nie, Y., Long, C., Zhang, Q., Li, G.: A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In: ICCV. pp. 13588–13597 (2021)
24. Luo, W., Liu, W., Gao, S.: A revisit of sparse coding based anomaly detection in stacked RNN framework. In: ICCV. pp. 341–349 (2017)
25. Luo, W., Liu, W., Gao, S.: Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection. Neurocomputing **444**, 332–337 (2021)
26. Markovitz, A., Sharir, G., Friedman, I., Zelnik-Manor, L., Avidan, S.: Graph embedded pose clustering for anomaly detection. In: CVPR. pp. 10539–10547 (2020)
27. Micorek, J., Possegger, H., Narnhofer, D., Bischof, H., Kozinski, M.: MULDE: Multiscale log-density estimation via denoising score matching for video anomaly detection. In: CVPR. pp. 18868–18877 (2024)
28. Morais, R., Le, V., Tran, T., Saha, B., Mansour, M., Venkatesh, S.: Learning regularity in skeleton trajectories for anomaly detection in videos. In: CVPR. pp. 11996–12004 (2019)
29. Noghre, G.A., Pazho, A.D., Tabkhi, H.: An exploratory study on human-centric video anomaly detection through variational autoencoders and trajectory prediction. In: WACV. pp. 995–1004 (2024)
30. Piciarelli, C., Micheloni, C., Foresti, G.L.: Trajectory-based anomalous event detection. IEEE TCSVT **18**(11), 1544–1554 (2008)
31. Rodrigues, R., Bhargava, N., Velmurugan, R., Chaudhuri, S.: Multi-timescale trajectory prediction for abnormal human activity detection. In: WACV. pp. 2626–2634 (2020)
32. Singh, A., Jones, M.J., Learned-Miller, E.G.: EVAL: Explainable video anomaly localization. In: CVPR. pp. 18717–18726 (2023)
33. Singh, A., Jones, M.J., Learned-Miller, E.G.: Tracklet-based explainable video anomaly localization. In: CVPR. pp. 3992–4001 (2024)
34. Song, I., Joo, M., Kwon, J., Jeon, E., Lee, J.: Instance-aligned captions for explainable video anomaly detection. arXiv preprint arXiv:2601.08155 (2026)
35. Song, I., Joo, M., Kwon, J., Lee, J.: Video question answering for people with visual impairments using an egocentric 360-degree camera. arXiv preprint arXiv:2405.19794 (2024)
36. Song, I., Lee, J.: Real-time traffic accident anticipation with feature reuse. In: ICIP (2025)
37. Song, I., Lee, S., Joo, M., Lee, J.: Anomaly detection for people with visual impairments using an egocentric 360-degree camera. In: WACV. pp. 2828–2837 (2025)
38. Stergiou, A., De Weerdt, B., Deligiannis, N.: Holistic representation learning for multitask trajectory anomaly detection. In: WACV. pp. 6729–6739 (2024)
39. Tudor Ionescu, R., Smeureanu, S., Alexe, B., Popescu, M.: Unmasking the abnormal events in video. In: ICCV. pp. 2895–2903 (2017)

40. Wang, G., Wang, Y., Qin, J., Zhang, D., Bao, X., Huang, D.: Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In: ECCV. pp. 494–511 (2022)
41. Wu, C., Shao, S., Tunc, C., Satam, P., Hariri, S.: An explainable and efficient deep learning framework for video anomaly detection. Cluster Computing **25**(4), 2715–2737 (2022)
42. Wu, R., Chen, Y., Xiao, J., Li, B., Fan, J., Dufaux, F., Zhu, C., Liu, Y.: DA-flow: Dual attention normalizing flow for skeleton-based video anomaly detection. IEEE TMM (2025)
43. Yan, C., Zhang, S., Liu, Y., Pang, G., Wang, W.: Feature prediction diffusion model for video anomaly detection. In: ICCV. pp. 5527–5537 (2023)
44. Yu, G., Wang, S., Cai, Z., Zhu, E., Xu, C., Yin, J., Kloft, M.: Cloze test helps: Effective video anomaly detection via learning to complete video events. In: ACM MM. pp. 583–591 (2020)
45. Zaheer, M.Z., Mahmood, A., Khan, M.H., Segu, M., Yu, F., Lee, S.I.: Generative cooperative learning for unsupervised video anomaly detection. In: CVPR. pp. 14744–14754 (2022)
46. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: ByteTrack: Multi-object tracking by associating every detection box. In: ECCV. pp. 1–21 (2022)
47. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: ICCV. pp. 3702–3712 (2019)
48. Zhu, L., Wang, L., Raj, A., Gedeon, T., Chen, C.: Advancing video anomaly detection: A concise review and a new dataset. NeurIPS **37**, 89943–89977 (2024)