Holistic Representation Learning for Multitask Trajectory Anomaly Detection

Alexandros Stergiou   Brent De Weerdt   Nikos Deligiannis

Vrije Universiteit Brussel, Belgium & imec, Belgium

`<first>.<last>@vub.be`

Abstract

Video anomaly detection deals with the recognition of abnormal events in videos. Apart from the visual signal, video anomaly detection has also been addressed with the use of skeleton sequences. We propose a holistic representation of skeleton trajectories to learn expected motions across segments at different times. Our approach uses multitask learning to reconstruct any continuous unobserved temporal segment of the trajectory allowing the extrapolation of past or future segments and the interpolation of in-between segments. We use an end-to-end attention-based encoder-decoder. We encode temporally occluded trajectories, jointly learn latent representations of the occluded segments, and reconstruct trajectories based on expected motions across different temporal segments. Extensive experiments on three trajectory-based video anomaly detection datasets show the advantages and effectiveness of our approach with state-of-the-art results on anomaly detection in skeleton trajectories[†].

1. Introduction

Everyday life is full of expected behaviors. Examples include; walking towards a destination, sitting down, or riding a bicycle. Such actions and behaviors are defined by the predictability of their motions that are part of an expected sequence. In computer vision, motion can be represented by trajectories, segments of which include *past*, *present*, and *future* events. Anomalies can occur at different times within a trajectory and are usually perceived through abrupt or sudden motion changes. The resulting anomalous trajectory segments deviate from the expected trajectory path, as illustrated in Figure 1.

Video anomaly detection (VAD) is the open-set task of detecting anomalous actions and motions by learning robust representations of expected behaviors in normal events. The detection of abnormal events in videos has gained traction with applications in surveillance [37, 49], social net-
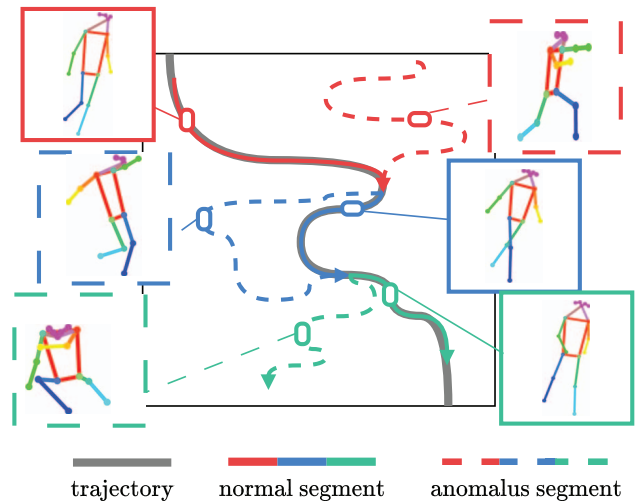
Figure 1. **Holistic representation of trajectory segments**. Given a normal trajectory in grey, past segments (Pst) are extrapolated based on future segments, present segments (Prs) are interpolated from proceeding and succeeding segments, and future segments (Ftr) are extrapolated from past segments. During training, reconstructed trajectories are compared to ground truth trajectories. In inference, the reconstruction error of task-specific segments is used to distinguish normal and anomalous trajectories.

works [46], and monitoring of daily activities [41]. The majority of trajectory-based VAD methods [16, 17, 35, 36] are trained to only extrapolate the future of normal trajectories. Abnormalities are detected at inference from the reconstruction error between predicted expected trajectories and the ground truth. Due to the continuous nature of trajectories, their reconstruction at any point in time is challenging in real-world scenarios as the future or past may not be observed and occlusions or pose estimation errors may lead to partial observations of the present.

We introduce a unifying framework that jointly addresses these challenges and enables the representation and reconstruction of expected normal motions at different trajectory segments. We use a holistic representation of skeleton trajectories as a composition of past, present, and future segments. We use an encoder-decoder [24] to first en-

code partial observations of normal trajectories with different temporal segments being occluded. Latent representations of the occluded segments are learned regressively through positive and negative pairs drawn from full trajectories. The learned representations are then decoded back to the input space and compared to ground truth segments. During training, our model learns to encapsulate information relating to normal motions and inter/extrapolates the occluded trajectory segments. At inference, the model predicts a corresponding normal trajectory that is compared to the ground truth trajectory of the unseen abnormal scene. Our method's simplicity allows the detection of abnormalities across multiple temporal locations within the trajectory and the use of trajectories with different temporal lengths.

In summary, our contributions are as follows: (i) We propose a holistic representation of trajectories for VAD to detect anomalies in past, present, or future trajectory segments. (ii) We train an encoder-decoder model with multitask learning to jointly learn latent representations of occluded past, present, and future trajectory segments and reconstruct them back to the input space. (iii) We evaluate our approach on three skeleton VAD datasets: HR-ShanghaiTech Campus (HR-STC) [32,36], HR-Avenue [30, 36], and (HR-)UBnormal [1, 17]. We consistently outperform prior work on extrapolating the future and we additionally provide baselines on the remaining two tasks.

2. Related work

Skeleton trajectories modeling. Motion modeling in videos with skeleton trajectories has been widely studied for tasks such as action recognition [9, 19, 42, 48], person re-identification [10, 11], human-computer interaction [26], virtual reality [7], and robotics [3]. A large portion of skeleton-based approaches has relied on graph convolutional networks [4–6,21,56] that model sequences as spatiotemporal graphs. Recent approaches have also been based on the representation of trajectories as spatiotemporal volumes through heatmaps [9]. Utilizing the continuity of time to create a singular signal for trajectories has also recently gained traction [50, 51] as it provides a flexible representation of the progression of actions. We adopt a similar approach by minimizing the reconstruction errors between the ground truth trajectory and the predicted trajectory reconstructed by the decoder. Jointly, the correlation of segments at different times and their continuous trajectory is maximized with our model learning to encapsulate trajectory information across different segments.

Video anomaly detection. The majority of anomaly detection methods in the literature [2, 13, 16, 17, 23, 28, 31, 32, 35, 36, 38, 44, 45, 49, 52, 54, 58, 59] are autoregressive, where abnormal behavior is only defined by inferring future skeletons from previous ones. However, these approaches are incapable of modeling the continuity of tra-

jectories in full and can only reconstruct a future segment of the trajectory. Some methods use a score function to maximize the embedding space distance between normal and abnormal instances [13, 38, 45, 52, 54, 58]. A different line of approaches learn to reconstruct normal trajectory segments with anomalous segments only present in inference [2, 23, 28, 31, 32, 44, 59]. Morais *et al.* [36] used a recurrent encoder-decoder to predict and reconstruct proceeding frames. Following works have included clustering embeddings of similar classes [35] or hyperspherical representations of latents [17]. Luo *et al.* [33] represented trajectories as graphs and used Graph Convolutions for predicting future keypoints. Other approaches have jointly studied the tasks of learning the arrow of time, predicting motion irregularity, and predicting object appearance [20]. Flaborea *et al.* [16] used a diffusion-based model to synthesize future poses conditioned on past trajectories. Despite strong baseline results, these approaches are not capable of learning to discriminate between trajectories of normal and abnormal behavior beyond the temporal limits of future segments.

Self supervised representations. Learning temporal signal representations through contrastive learning has been explored for video [8, 14, 39, 53], audio [15, 18, 55], and pose [25, 47, 60]. For trajectories, works have focused on using augmentations comparing positive and negative instances [22, 27, 29, 34]. A set of recent approaches has also used self-supervision to represent hierarchical relationships between segments [40, 57]. Surís and Vondrick [51] have represented segments through a probability distribution in a latent space. In our work, we use self-supervision to correlate and compare representations of trajectory segments and learned latent representations of normal trajectories.

3. Method

In this section, we overview our approach, shown in Figure 2. We first introduce our holistic multitask approach for learning normality in trajectories in Section 3.2. Each latent space trajectory segment encodes the expected progression of a normal trajectory at different times through positive and negative-pair self-supervision defined in Section 3.3. Latent segment representations are decoded to the input space and compared to normal/abnormal trajectories as detailed in Section 3.4.

3.1. Definitions

We use standard VAD definitions from the literature [16, 17, 35, 36] where trajectories correspond to anomalous events characterized by irregular motions. Trajectories of skeleton keypoints from non-anomalous instances are used to learn normal behaviors, which are then reconstructed and compared to abnormal sequences at inference. We denote the full continuous trajectory over T frames as $\mathbf{v} \in \mathbb{R}^{T \times N}$, where N are the spatial coordinates of each point $\mathbf{x}$ over
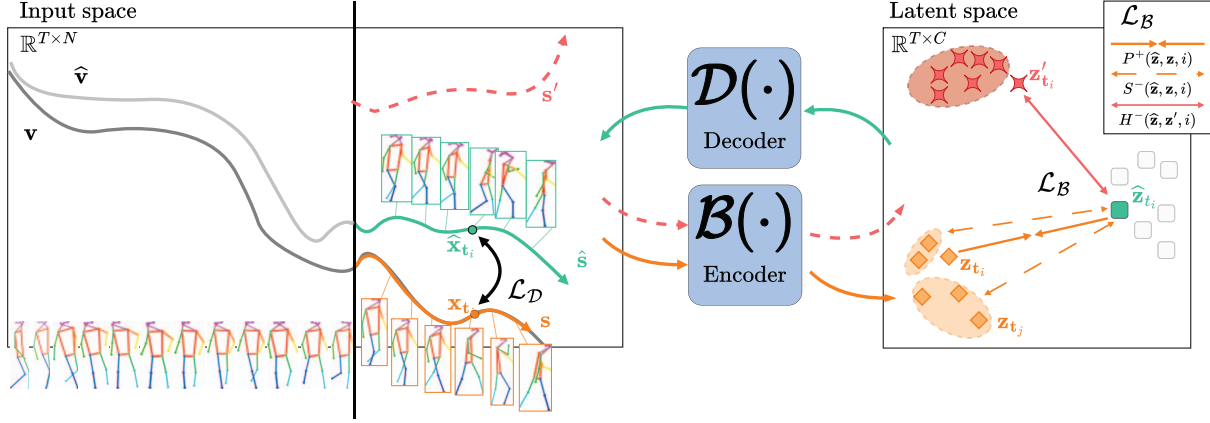
Figure 2. **Method overview**. Segment $\mathbf{s}$ is occluded from trajectory $\mathbf{v} \in \mathbb{R}^{T \times N}$. Each spatial point $\mathbf{x}$ at time t_i is mapped to $\mathbf{z}_{t_i}$ in the latent space $\mathbb{R}^{T \times C}$ by encoder $\mathcal{B}$. Tensor $\widehat{\mathbf{z}}$ corresponding to expected latent representations is learned contrastively from positive pair $(\widehat{\mathbf{z}}_{t_i}, \mathbf{z}_{t_i})$, soft negative pair $(\widehat{\mathbf{z}}_{t_i}, \mathbf{z}_{t_j})$ for $t_j \neq t_i$, and hard negative pair $(\widehat{\mathbf{z}}_{t_i}, \mathbf{z}'_{t_i})$. Latent $\mathbf{z}_{t_i}$ of point $\mathbf{x}$ at time t_i is used as a positive pair to $\widehat{\mathbf{z}}_{t_i}$. The remaining projections $\mathbf{z}_{t_j}$ from segment $\mathbf{s}$ are soft negative pairs to $\widehat{\mathbf{z}}_{t_i}$. Latents $\mathbf{z}_{t_i}$ of segments $\mathbf{s}'$ from other trajectories are considered hard negatives. The learned $\widehat{\mathbf{z}}_{t_i}$ is projected back to point $\widehat{\mathbf{x}}$ in the input space $\mathbb{R}^{T \times N}$ through decoder $\mathcal{D}$. Projected points $\widehat{\mathbf{x}}$ in the estimated trajectory $\widehat{\mathbf{v}}$ are compared to their corresponding points $\mathbf{x}$ in $\mathbf{v}$.

$\mathbf{t} = \{1, ..., T\}$ locations. We define a sub-sequence as continuous segment $\mathbf{s}$ over $\widehat{\mathbf{t}}$ temporal locations.

The motion information of the entire trajectory $\mathbf{v}$ is encoded from the input space into representation $\mathbf{z}$ within latent space $\mathbb{R}^{T \times C}$. Each point $\mathbf{z}_{t_i}$ in the latent space corresponds to a spatial point of the trajectory at a time $t_i \in \mathbf{t}$. Our goal is to approximate the occluded segment $\mathbf{s}$ and its representations when only part of the trajectory is observed $\mathbf{v} \setminus \mathbf{s} = \{\mathbf{x} : \mathbf{x} \in \mathbf{v} \text{ and } \mathbf{x} \notin \mathbf{s}\}$. We use an encoder $\mathcal{B}$ to encode each point $\mathbf{x}$ at t_i into the latent space $\mathcal{B}(\{\mathbf{x}; \mathbf{v} \setminus \mathbf{s}\})$. The latent representations are then combined with a learned tensor to form an estimated latent trajectory. We use a decoder $\mathcal{D}$ to decode each representation at temporal point t_i back to the input space and obtain point $\widehat{\mathbf{x}}$.

3.2. Multitask Holistic Trajectories

Typically, a natural choice for the detection of anomalies would be to extrapolate expected motion patterns only for future segments. However, we argue that jointly learning multiple trajectory segments is crucial for both distinguishing anomalies that may occur at different times as well as learning a high-level understanding of the global trajectory. For instance, the future trajectory of a person who is slowing down is less likely to include abnormalities than the past. Additionally, the continuity of trajectories relies on pose tracking with occlusions at both the keypoint and frame levels present at different times. We thus propose a holistic representation of trajectories for past, present, and future segments. Given partial trajectory $\mathbf{v} \setminus \mathbf{s}$ we predict occluded segments $\mathbf{s} = \{\mathbf{x}_{t_i} : t_i \in \widehat{\mathbf{t}}\}$ for three tasks.

Predicting the future given the past (Ftr). *Future segments* over $\widehat{\mathbf{t}} = \{T_{Ftr}, ..., T\}$ are estimated from partial tra-

jectory $\mathbf{v} \setminus \mathbf{s}$ composed of only past segments.

Predicting the past given the future (Pst). *Past segments* over $\widehat{\mathbf{t}} = \{1, ..., T_{Pst}\}$ are estimated from partial trajectory $\mathbf{v} \setminus \mathbf{s}$ of only future segments.

Predicting the present given both past and future (Prs). In-between segments over $\widehat{\mathbf{t}} = \{T_{Pst} + 1, ..., T_{Ftr} - 1\}$, are referred to as *present segments* and estimated from partial trajectory $\mathbf{v} \setminus \mathbf{s}$ composed of both past and future segments.

We explore these tasks jointly in a multitask training scheme. Latent representations for each of the occluded Pst, Prs, and Ftr, segments are learned jointly and reconstructed back to the input space by the decoder.

3.3. Latent Representation Leaning

We use an attention-based encoder $\mathcal{B}$ applied only on the un-occluded part of trajectories $\mathbf{v} \setminus \mathbf{s}$. Similar to [36] we use the trajectories of both skeleton keypoints and the corresponding bounding box corners. We define a learnable tensor $\mathbf{u} = \{\widehat{\mathbf{z}}_{t_i} : t_i \in \mathbf{t}\}$ of equal size as the full trajectory's representations $|\mathbf{u}| = |\mathcal{B}(\mathbf{v})|$. For each task, we select segment $\mathbf{u}_s = \{\widehat{\mathbf{z}}_{t_i} : t_i \in \widehat{\mathbf{t}}\}$ corresponding to the occluded $\mathbf{s}$ in $\mathbf{v}$. Tensor segment $\mathbf{u}_s$ is combined with the observed trajectory latents:

$$\mathbf{a} = \Phi(\mathcal{B}(\{\mathbf{x}; \mathbf{v} \setminus \mathbf{s}\}) \cup \mathbf{u}_s) \tag{1}$$

where Φ is a function that reorders $\mathcal{B}(\{\mathbf{x}; \mathbf{v} \setminus \mathbf{s}\}) \cup \mathbf{u}_s$ given the temporal location of $\mathbf{s}$ within $\mathbf{v}$. We run inference to also obtain the representations of the entire sequence $\mathbf{z} = \mathcal{B}(\mathbf{v})$.

Consequently, we want each of the learned vectors $\widehat{\mathbf{z}}_{t_i} \in \mathbf{u}_s$ to be drawn closer to the corresponding trajectory representations $\mathbf{z}_{t_i} \in \mathcal{B}(\mathbf{v})$ at temporal point t_i while being pushed away from representations at other trajectory points

6717

or other random points. To encourage this, we define an objective based on latent space positive and negative pairs.

Positive pairs. We wish to minimize the distance between learned vector $\widehat{\mathbf{z}}_{t_i}$ and the trajectory representation $\mathbf{z}_{t_i}$ at t_i. We thus define pair $P^+(\widehat{\mathbf{z}}, \mathbf{z}, i) = \|\widehat{\mathbf{z}}_{t_i} - \mathbf{z}_{t_i}\|_2$ as a *positive*.

Soft negative pairs. Given the remaining representations $\mathbf{z}_{t_j}$ at temporal locations $t_j \neq t_i$ we want to maximize their distance to $\widehat{\mathbf{z}}_{t_i}$. However, as $\mathbf{z}_{t_j}$ belongs to the same trajectory for which $\widehat{\mathbf{z}}_{t_i}$ is an estimate of, they are bound to include some similarities. We thus treat $(\widehat{\mathbf{z}}_{t_i}, \mathbf{z}_{t_j})$ as *soft negative* pairs for which their difference is regularized by their temporal distance. We want to penalize less pairs of points that are temporally close comparatively to pairs of points further away: $\mathcal{R}(t_i, t_{i+1}) < \mathcal{R}(t_i, t_{i+2})$. The soft negative pairs penalty is defined as:

$$S^-(\widehat{\mathbf{z}}, \mathbf{z}, i) = \sum_{t_j \in \widehat{\mathbf{t}}} \mathcal{R}(t_i, t_j) \|\widehat{\mathbf{z}}_{t_i} - \mathbf{z}_{t_j}\|_2, \text{ where}$$

$$\mathcal{R}(t_i, t_j) = \frac{\beta \|t_i - t_j\|_1}{\max(\|t_i - t_k\|_1 : t_k \in \widehat{\mathbf{t}})} \tag{2}$$

where β is a hyperparameter used to adjust the penalization.

Hard negative pairs. We select latents of segments $\mathbf{s}'$ from other trajectories as *hard negative* pairs. As each segment $\mathbf{s}$ is unique to trajectory $\mathbf{v}$, no representation $\mathbf{z}'$ from segment $\mathbf{s}'$ should correspond to $\widehat{\mathbf{z}}_{t_i}$. Thus, we also maximize their distance $H^-(\widehat{\mathbf{z}}, \mathbf{z}', i) = \|\widehat{\mathbf{z}}_{t_i} - \mathbf{z}'_{t_i}\|_2$.

We use a self-supervised triplet-loss $\mathcal{L}_\mathcal{B}$ with an additional *soft negative* pair term. This forces encoder $\mathcal{B}$ to attain estimated representations $\widehat{\mathbf{z}}_{t_i}$ close to latents $\mathbf{z}_{t_i}$ at time t_i. It also reduces the resemblance to remaining representations $\mathbf{z}_{t_j}$ from other temporal locations as well as representations $\mathbf{z}'_{t_i}$ from segments $\mathbf{s}'$ of other trajectories.

$$\mathcal{L}_\mathcal{B} = \sum_{t_i \in \widehat{\mathbf{t}}} \max(P^+(\widehat{\mathbf{z}}, \mathbf{z}, i) - S^-(\widehat{\mathbf{z}}, \mathbf{z}, i) \\ - H^-(\widehat{\mathbf{z}}, \mathbf{z}', i) + \gamma, 0) \tag{3}$$

where γ is a margin hyperparameter.

3.4. Trajectory segment reconstruction

We wish to decode the representations of the trajectory's spatial points from latent space $\mathbb{R}^{T \times C}$ back into the input space $\mathbb{R}^{T \times N}$. Decoder $\mathcal{D}$ takes $\mathbf{a}$ from (1) and projects representations $\mathcal{B}(\{\mathbf{x}; \mathbf{v} \setminus \mathbf{s}\})$ of the un-occluded trajectory $\mathbf{v} \setminus \mathbf{s}$ alongside the occluded learned segment $\mathbf{u}_s$, back to $\mathbb{R}^{T \times N}$. As both ground truth trajectory $\mathbf{v}$ and estimated $\widehat{\mathbf{v}} = \mathcal{D}(\mathbf{a})$ are available, the decoder is explicitly trained on reconstruction and not extra/interpolation.

$$\mathcal{L}_\mathcal{D} = \frac{1}{T} \sum_{t_i \in \mathbf{t}} \|\widehat{\mathbf{x}}_{t_i} - \mathbf{x}_{t_i}\|_2 \tag{4}$$

where $\widehat{\mathbf{x}} \in \widehat{\mathbf{v}}$ and $\mathbf{x} \in \mathbf{v}$. The decoder is trained to regress the error between reconstructed $\widehat{\mathbf{x}}$ and ground truth $\mathbf{x}$ at temporal location t_i in the input space $\mathbb{R}^{T \times N}$. At inference, only decoded points $\widehat{\mathbf{x}} \in \widehat{\mathbf{s}}$ are compared to points $\mathbf{x} \in \mathbf{s}$.

We combine the encoder and decoder losses from (3) and (4), and define our multitask learning objective:

$$\mathcal{L} = \mathcal{L}_\mathcal{B} + \lambda \mathcal{L}_\mathcal{D} \tag{5}$$

where λ is a hyperparameter that is tuned according to the trajectory belonging to either a skeleton keypoint or a bounding box corner. At each training step, we extrapolate/interpolate an equal number of trajectory segments for each Pst, Prs, and Ftr task before a gradient update. As each task uses a different segment $\mathbf{u}_s$, updates are computed accumulatively for the entirety of $\mathbf{u}$ to stabilize training.

4. Experiments

The datasets used, alongside implementation and training details are described in Section 4.1. We compare our proposed approach to state-of-the-art models in Section 4.2. We show qualitative results on all three tasks in Section 4.3 followed by ablation studies in Section 4.4.

4.1. Experimental details

Datasets. We report our framework's performance over a diverse set of VAD datasets. Human-Related ShanghaiTech Campus (*HR-STC*) [32, 36] consists of 13 scenes with 303 train and 101 test videos containing 130 anomalous events. *HR-Avenue* [30, 36] includes a single scene with 16 train and 21 test videos. *UBnormal* [1] is a collection of 29 synthetic scenes synthesized with Cinema4D and natural backgrounds with 186 normal train and 211 test videos. *HR-UBnormal* [1, 17] is a subset of UBnormal introduced by [17] that includes only HR anomalies in the test set. For all datasets, we use the provided skeleton keypoints and bounding box coordinates detected with AlphaPose [12]. Following works in the literature [1, 16, 17, 28, 30, 35, 36] we use the Area Under the Receiver Operating Characteristic Curve (AUC) score as our performance metric.

Model settings. We train our model jointly on all three tasks; Ftr, Pst, and Pst. We directly compare to the state-of-the-art for Ftr as it has been primarily used for VAD, and provide baselines for Pst, and Pst. Overall, we employ two STTFormers [43], one as the encoder $\mathcal{B}$ for projecting trajectory points to $\mathbb{R}^{T \times C}$ and the second is inversed and used as the decoder to project points back to $\mathbb{R}^{T \times N}$. Unless otherwise specified, we use $T = 18$ trajectory frames with segment lengths of $|\widehat{\mathbf{t}}| = 6$. This corresponds to input sequences of $18 \times (17 + 4) \times 2$, where 17 is the number of skeleton joints, 4 is the corners of the bounding box of the skeleton, and 2 corresponds to their spatial locations. The size of the learned latent tensor $\mathbf{u}$ is $|\mathbf{u}| = 18 \times 256$. We

Table 1. **AUC performance of VAD methods on HR-STC, HR-Avenue, HR-UBnormal, and UBnormal**. Ftr is the common task addressed by all VAD methods in the literature. Top performances for each task and for each dataset are in **bold**.

	HR-STC			HR-Avenue			HR-UBnormal			UBnormal		
	Ftr	Prs	Pst	Ftr	Prs	Pst	Ftr	Prs	Pst	Ftr	Prs	Pst
GEPC [35]	74.8	-	-	58.1	-	-	55.2	-	-	53.4	-	-
MPED-RNN [36]	75.4	69.8*	72.1*	86.2	81.4*	83.8*	61.2	59.3*	61.1*	60.6	58.5*	60.1*
COSKAD [17]	77.1	-	-	87.3	-	-	65.5	-	-	65.0	-	-
MoCoDAD [16]	77.6	-	-	89.0	-	-	**68.4**	-	-	**68.3**	-	-
TrajREC (ours)	**77.9**	**73.5**	**75.7**	**89.4**	**86.3**	**87.6**	68.2	**64.1**	**66.8**	68.0	**63.6**	**66.4**

train with $1e^{-4}$ base learning rate and a batch size of 512. We use $\beta = 0.001$, and $\gamma = 0.1$, with $\lambda = 5$ for joint and $\lambda = 3$ for bounding box corner trajectories. An overview of AUC scores with different hyperparameter combinations is shown in §S1 in the supplementary material.

4.2. Comparative results

In Table 1 we report AUC scores on the three tasks; extrapolation of future keypoints from past keypoints Ftr, extrapolation of past keypoints from future keypoints Pst, and interpolation of in-between/present keypoints from both future and past keypoints Prs. Current models are not capable of jointly extra/interpolating keypoints and are instead only limited to future extrapolation Ftr. We select MPED-RNN [36] as a baseline for the previously unexplored past extrapolation Pst and present interpolation Prs, due to its popular use as a VAD baseline and open-source implementation. The design of MPED-RNN does not allow multitask learning so we retrain individually for Pst and Prs.

HR-STC. Across the Ftr task we show that our proposed approach outperforms all other single-task approaches. We achieve a +2.5 AUC score improvement over the baseline [36] and +0.3 over the best-performing single-task model [16]. For Pst our model achieves a 75.7 AUC score outperforming [36] by +3.6. Similar score increases are also observed for the Prs task with a +3.7 improvement over the baseline. Our method's flexibility enables learning and inferring holistic representations of the entire trajectory which can evidently benefit all three tasks jointly.

HR-Avenue. Compared to the baseline, we observe a +3.2, +3.8, and +4.9 improvement in the AUC score for the Ftr, Pst, and Prs tasks respectively. Our method also surpasses the previous state-of-the-art models [16] on the Ftr task with a +0.4 increase in the AUC score.

HR-UBnormal. We observe that the relative increase in performance also varies across datasets with +7.0 for the Ftr, +4.8 for Prs, and +5.7 for Pst AUC score increase over the baseline. We believe this to be due to the scene diversity of the dataset. Due to UBnormal using synthetic data, we speculate that methods based on generative models such as [16] can better model motions in synthetic trajectories.

Table 2. **AUC scores on HR-STC and HR-Avenue over different trajectory lengths**. In all settings, the predicted segments are equal to a third of the trajectory's length.

	HR-STC			HR-Avenue		
	Ftr	Prs	Pst	Ftr	Prs	Pst
Trajectory length $T = 18$, Segment length $|\hat{t}| = 6$						
Baseline [36]	75.4	69.8	72.1	86.2	81.4	83.8
TrajREC (ours)	**77.9**	**73.5**	**75.7**	**89.4**	**86.3**	**87.6**
Trajectory length $T = 24$, Segment length $|\hat{t}| = 8$						
Baseline [36]	72.9	67.2	69.8	84.0	80.8	82.6
TrajREC (ours)	**76.4**	**72.7**	**74.7**	**88.5**	**85.7**	**86.3**
Trajectory length $T = 36$, Segment length $|\hat{t}| = 12$						
Baseline [36]	69.2	66.3	67.4	81.1	78.6	80.1
TrajREC (ours)	**75.3**	**72.5**	**73.3**	**86.8**	**84.1**	**84.7**

UBnormal. The results on the full UBnormal dataset follow the same trends as those from the HR-UBnormal subset. Relative to the baseline, the largest improvement across tasks is observed for Ftr with a +7.4 AUC score increase followed by Pst with +6.3 improvement. For the Prs task we note a +5.1 improvement in the AUC score.

4.3. Qualitative results

In Figure 3 we demonstrate examples of reconstructed trajectory segments for normal and anomalous sequences. Extrapolated and interpolated segments of normal trajectories for Ftr, Prs, and Pst are shown in Figures 3a, 3b, and 3c. For all tasks, the model is capable of learning continual sequences with the motions in reconstructed skeletons progressing smoothly. In instances where not all joints are visible or recognized in the input (Figure 3c) the spatial locations of the predicted joints are reasonable in relation to the neighboring observed keypoints. In Figures 3d, 3e, and 3f we show reconstructions for segments that include anomalies. As the behaviors and motions of these segments are not part of the training set, extrapolated or interpolated segments can be easily detected as anomalous based on the high reconstruction error. In extrapolation tasks, the reconstructed trajectory segments resemble motions of expected normal behavior. In contrast, interpolation requires infilling
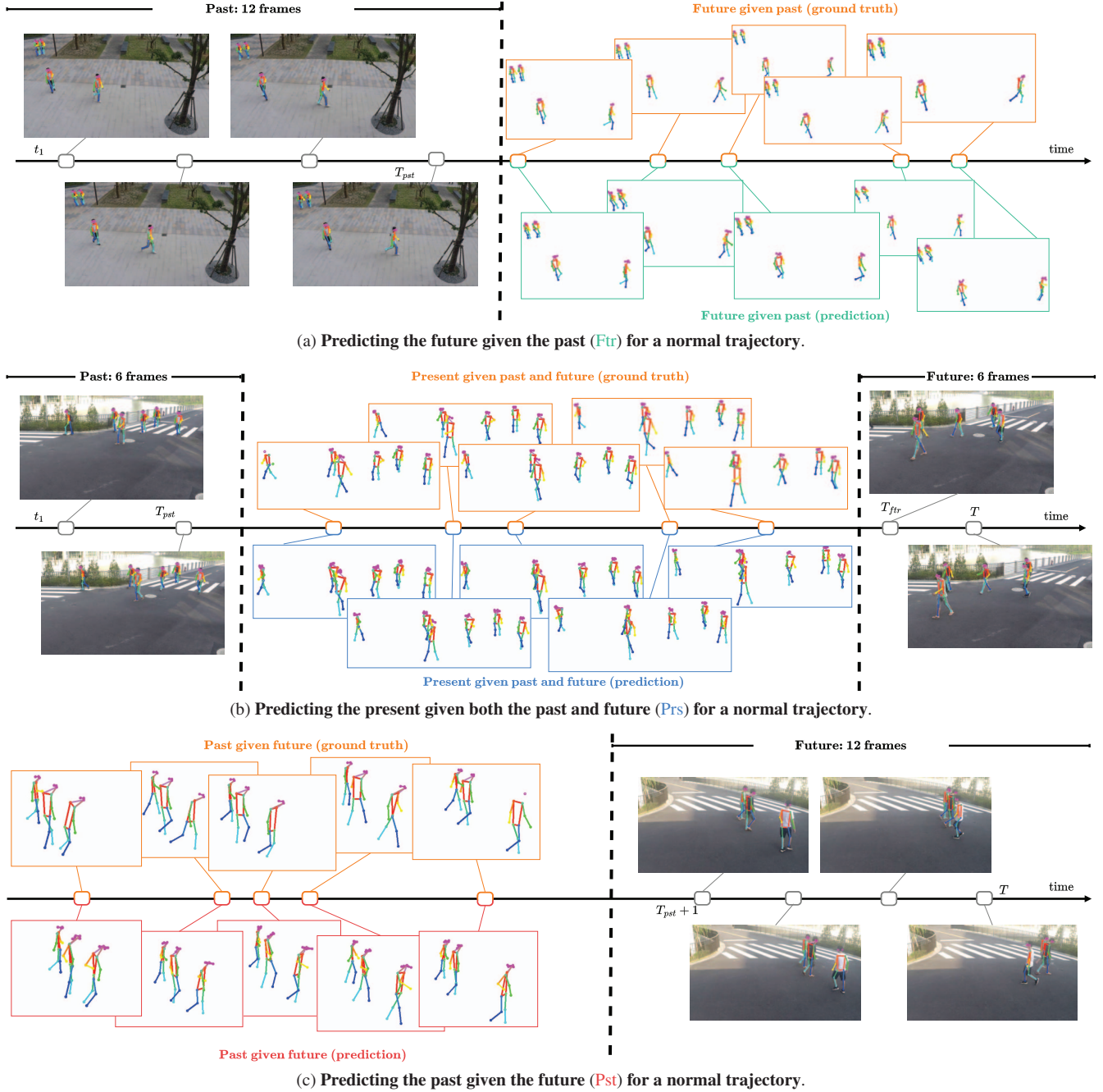
---

*Inhouse retrained for the specific task.

(a) **Predicting the future given the past** (Ftr) **for a normal trajectory**.

(b) **Predicting the present given both the past and future** (Prs) **for a normal trajectory**.

(c) **Predicting the past given the future** (Pst) **for a normal trajectory**.

Figure 3. **Normal and abnormal skeleton trajectory reconstruction for HR-STC**. Input trajectories are keypoints from 18 frames with each of the predicted segments for Ftr, Prs, and Pst task being 6 frames. Both extrapolations of future and past segments as well as the interpolation of in-between segments are sensible predictions of normal behaviors across all skeletons in each scene.

of occluded segments based on the given past and future keypoints. As proceeding and succeeding frames may also be anomalous (Figure 3e), the reconstructions produced are of low confidence, showing that no normal segment can exist when the frames at the start and end are anomalous. Additional examples are available in §S2.

4.4. Ablation studies

In this section, we ablate over trajectory lengths reporting model performance on segments of increased durations. Additionally, we evaluate our method over different learning schemes and single-task settings. Finally, we present AUC scores over multiple runs with different initializations and show reconstruction examples for each run.
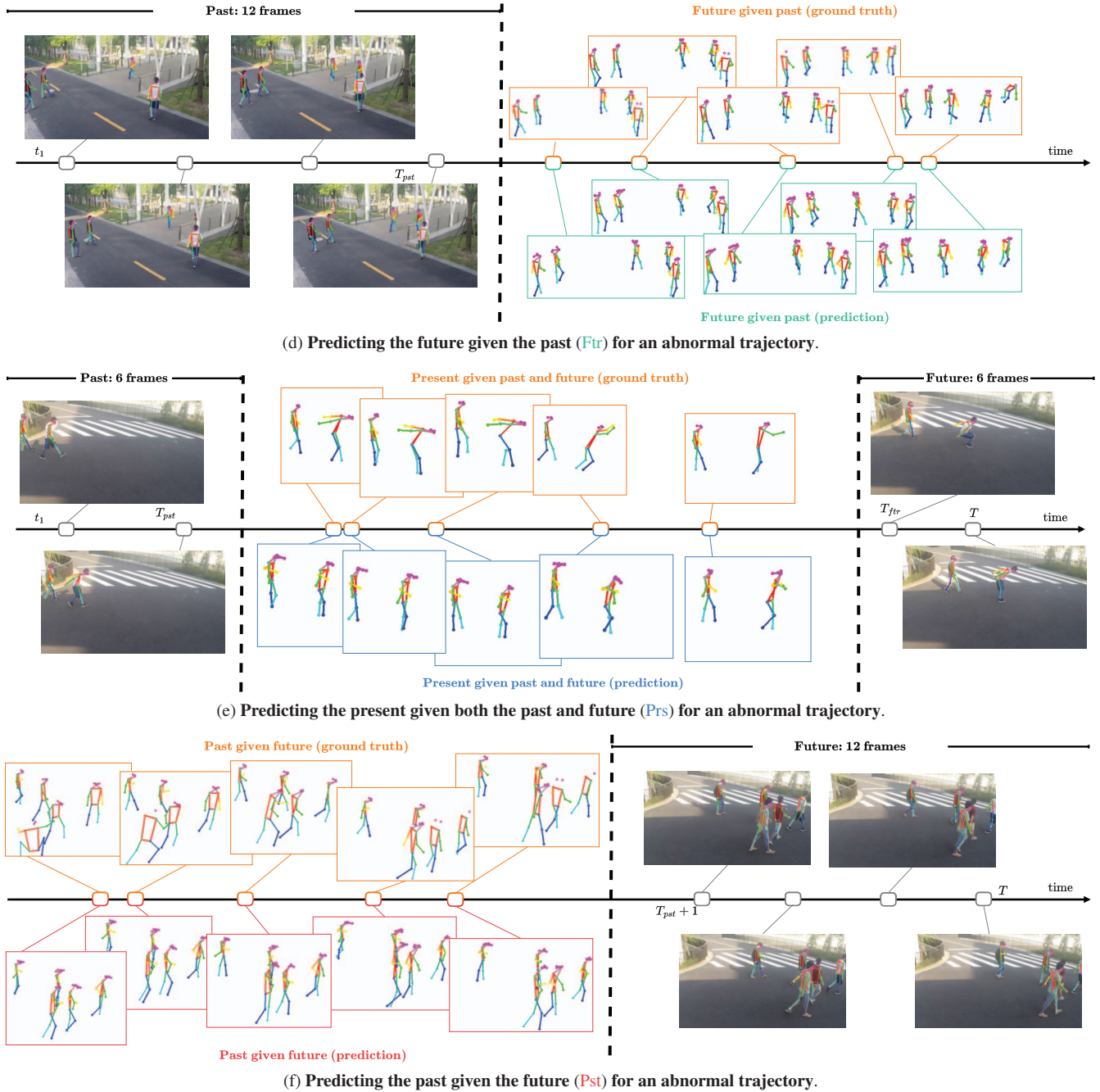
6720

(d) **Predicting the future given the past (Ftr) for an abnormal trajectory.**

(e) **Predicting the present given both the past and future (Prs) for an abnormal trajectory.**

(f) **Predicting the past given the future (Pst) for an abnormal trajectory.**

Figure 3. **Normal and abnormal skeleton trajectory reconstruction for HR-STC**. Input trajectories are keypoints from 18 frames with each of the predicted segments for Ftr, Prs, and Pst task being 6 frames. Both extrapolations of future and past segments as well as the interpolation of in-between segments are sensible predictions of normal behaviors across all skeletons in each scene (continued).

Trajectory and segment length. Trajectories and their corresponding segments can vary in length. We evaluate our method on HR-STC and HR-Avenue over multiple trajectory lengths $|T| \in \{18, 24, 36\}$ while keeping the size of the occluded segment equal to a third of the trajectory. As shown in Table 2, our model consistently outperforms the baseline [36] across tasks and trajectory lengths. An aver-

age -2.0 and -2.6 drop in the AUC scores across tasks is observed for HR-STC and HR-Avenue respectively when increasing $|T| = 18 \rightarrow |T| = 36$. This drop is more prevalent for the baseline with -4.8 and -3.9 for HR-STC and HR-Avenue. Thus, the holistic representation and multitask learning of trajectories is a better strategy for the reconstruction of segments from trajectories with longer durations.

6721

Table 3. **Ablations on HR-STC** with different learning scheme settings across different occluded segment $\hat{\mathbf{t}}$ sizes. Larger $|\hat{\mathbf{t}}|$ corresponds to smaller visible trajectories and reconstructed segments of larger lengths. The temporal length in all settings is $|T| = 18$.

$|\hat{\mathbf{t}}|$	Method	Ftr	Prs	Pst
3	Baseline [36]	76.1	70.4	72.2
	TrajREC w/o H^- (ours)	77.2	71.9	74.7
	TrajREC w/o S^- (ours)	77.5	73.0	75.6
	TrajREC single task (ours)	77.8	73.5	75.4
	TrajREC (ours)	**78.1**	**73.7**	**76.0**
6	Baseline [36]	75.4	69.8	72.1
	TrajREC w/o H^- (ours)	76.9	71.4	74.0
	TrajREC w/o S^- (ours)	77.2	71.8	75.2
	TrajREC single task (ours)	77.4	73.1	75.6
	TrajREC (ours)	**77.9**	**73.5**	**75.7**
9	Baseline [36]	72.9	69.1	71.7
	TrajREC w/o H^- (ours)	75.2	70.9	73.8
	TrajREC w/o S^- (ours)	75.6	71.8	74.5
	TrajREC single task (ours)	75.4	71.4	74.9
	TrajREC (ours)	**76.2**	**72.8**	**76.0**
12	Baseline [36]	70.6	66.8	69.4
	TrajREC w/o H^- (ours)	74.1	70.4	73.7
	TrajREC w/o S^- (ours)	74.7	71.3	74.2
	TrajREC single task (ours)	72.9	69.9	72.5
	TrajREC (ours)	**75.9**	**72.7**	**75.9**

Occlusion length. We compare the reconstruction performance of our approach at different occlusion lengths in Table 3. We maintain the temporal resolution of the input $|T| = 18$ and occlude segments of progressively increasing lengths; $|\hat{\mathbf{t}}| = \{3, 6, 9, 12\}$. Larger occlusions correspond to less of the input being observed. Models trained with multitask learning demonstrate only moderate decreases in their AUC score as the length of the occluded segments increases. We note that single-task models including both the baseline and our model trained individually per task $\mathbf{u}_s = \mathbf{u}$, show a notable performance decrease when the occluded segment's length is increased, as they do not learn a high-level understanding of the global trajectory.

Soft-hard negative pairs. We additionally include ablations on the effect of hard negative H^- and soft negative S^- pairs to the AUC score in Table 3. As shown, the removal of either H^- and S^- results in an overall decrease in AUC scores across tasks and occluded segment sizes. Marginally larger AUC score reductions are observed when H^- is removed which we believe is due to points from different trajectories being much stronger negative pairs than points from the current trajectory.

Multiple predictions. An important aspect of the reconstruction of normal behaviors corresponding to future, past, or present events is their predictability. VAD models need to robustly extrapolate (or interpolate) occluded trajectory segments of expected normal behaviors. In Table 4 we show AUC scores over multiple runs on HR-STC and HR-

Table 4. **AUC scores on HR-STC and HR-Avenue over multiple runs**. The best run is denoted with (best).

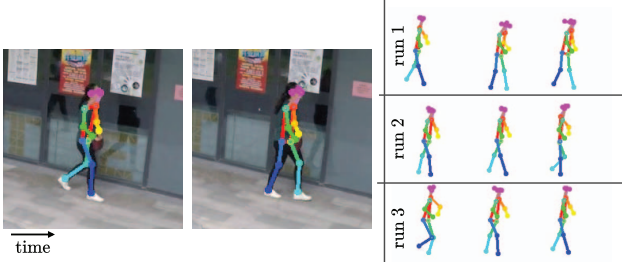|	HR-STC			HR-Avenue		
	Ftr	**Prs**	**Pst**	**Ftr**	**Prs**	**Pst**
run1 (best)	**77.9**	**73.5**	**75.7**	**89.4**	**86.3**	**87.6**
run2	77.6	73.4	75.2	89.4	86.0	87.1
run3	77.9	73.2	75.4	88.8	86.1	87.3

Figure 4. **Multiple Ftr on HR-STC**. Given the same input, we reconstruct future skeletons for each run corresponding to the model from Table 4. Predicted future joints vary across runs however all models learn robust representations of expected normal behavior.

Avenue, without any notable changes in the performance being observed. Examples of predicting future keypoints for each run from Table 4 are shown in Figure 4. Albeit their differences, all three of the reconstructed skeletons do correspond to expected normal motions.

5. Conclusions

We have proposed a holistic representation of trajectories with past, present, and future segments for VAD. Based on these segments, we introduce a multitask approach to capture normality in trajectories and jointly reconstruct past, present, and future segments. Temporally occluded trajectories are encoded and combined with a learned tensor. The latent representations of the occluded segments are learned through self-supervision with positive, soft-negative, and hard-negative pairs for each temporal location of the latent. Representations for the entire trajectory are decoded back to the input space and compared to the ground truth trajectory. Extensive experiments over three VAD datasets demonstrate the effectiveness of our approach. Additionally, we are the first to investigate the prediction of abnormal trajectories in the past as well as the present. We believe that this holistic representation of trajectories is a promising direction for future VAD research.

References

[1] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20143–20153. IEEE, 2022. 2, 4

[2] Samet Akçay, Amir Atapour-Abarghouei, and Toby P Breckon. Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. 2

[3] Chaitanya Bandi and Ulrike Thomas. Skeleton-based action recognition for human-robot interaction using self-attention mechanism. In *International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2021. 2

[4] Jinmiao Cai, Nianjuan Jiang, Xiaoguang Han, Kui Jia, and Jiangbo Lu. Jolo-gcn: mining joint-centered light-weight information for skeleton-based action recognition. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 2735–2744. IEEE, 2021. 2

[5] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *International Conference on Computer Vision (ICCV)*, pages 13359–13368. IEEE, 2021. 2

[6] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20186–20196. IEEE, 2022. 2

[7] UmurAybars Ciftci, Xing Zhang, and Lijun Tin. Partially occluded facial action recognition and interaction in virtual reality applications. In *International Conference on Multimedia and Expo (ICME)*, pages 715–720. IEEE, 2017. 2

[8] Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. Tclr: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding*, 219:103406, 2022. 2

[9] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2969–2978. IEEE, 2022. 2

[10] Amani Elaoud, Walid Barhoumi, Hassen Drira, and Ezzeddine Zagrouba. Analysis of skeletal shape trajectories for person re-identification. In *Advanced Concepts for Intelligent Vision Systems: 18th International Conference (ACIVS)*, pages 138–149. Springer, 2017. 2

[11] Amani Elaoud, Walid Barhoumi, Hassen Drira, and Ezzeddine Zagrouba. Modeling trajectories for 3d motion analysis. In *Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*, pages 409–429. Springer, 2020. 2

[12] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7157–7173, 2022. 4

[13] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14009–14018. IEEE, 2021. 2

[14] Basura Fernando and Samitha Herath. Anticipating human actions by correlating past with the future with jaccard similarity measures. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13224–13233. IEEE, 2021. 2

[15] Andres Ferraro, Xavier Favory, Konstantinos Drossos, Yuntae Kim, and Dmitry Bogdanov. Enriched music representations with multiple cross-modal contrastive learning. *IEEE Signal Processing Letters*, 28:733–737, 2021. 2

[16] Alessandro Flaborea, Luca Collorone, Guido Maria D'Amely di Melendugno, Stefano D'Arrigo, Bardh Prenkaj, and Fabio Galasso. Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection. In *International Conference on Computer Vision (ICCV)*, pages 10318–10329, 2023. 1, 2, 4, 5

[17] Alessandro Flaborea, Guido Maria D'Amely di Melendugno, Stefano D'arrigo, Marco Aurelio Sterpa, Alessio Sampieri, and Fabio Galasso. Contracting skeletal kinematic embeddings for anomaly detection. *arXiv preprint arXiv:2301.09489*, 2023. 1, 2, 4, 5

[18] Eduardo Fonseca, Diego Ortego, Kevin McGuinness, Noel E O'Connor, and Xavier Serra. Unsupervised contrastive learning of sound event representations. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 371–375. IEEE, 2021. 2

[19] Lin Geng Foo, Tianjiao Li, Hossein Rahmani, Qiuhong Ke, and Jun Liu. Unified pose sequence modeling. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13019–13030. IEEE, 2023. 2

[20] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12742–12752. IEEE, 2021. 2

[21] Pranay Gupta, Anirudh Thatipelli, Aditya Aggarwal, Shubh Maheshwari, Neel Trivedi, Sourav Das, and Ravi Kiran Sarvadevabhatla. Quo vadis, skeleton action recognition? *International Journal of Computer Vision*, 129(7):2097–2112, 2021. 2

[22] Marah Halawa, Olaf Hellwich, and Pia Bideau. Action-based contrastive learning for trajectory prediction. In *European Conference on Computer Vision (ECCV)*, pages 143–159. Springer, 2022. 2

[23] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 733–742. IEEE, 2016. 2

[24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009. IEEE, 2022. 1

[25] Seongyeong Lee, Hansoo Park, Dong Uk Kim, Jihyeon Kim, Muhammadjon Boboev, and Seungryul Baek. Image-free domain generalization via clip for 3d hand pose estimation. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 2934–2944. IEEE, 2023. 2

[26] Gang Li and Chunyu Li. Learning skeleton information for human action analysis using kinect. *Signal Processing: Image Communication*, 84:115814, 2020. 2

[27] Shuzhe Li, Wei Chen, Bingqi Yan, Zhen Li, Shunzhi Zhu, and Yanwei Yu. Self-supervised contrastive representation learning for large-scale trajectories. *Future Generation Computer Systems*, 2023. 2

[28] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6536–6545. IEEE, 2018. 2, 4

[29] Yuejiang Liu, Qi Yan, and Alexandre Alahi. Social nce: Contrastive learning of socially-aware motion representations. In *International Conference on Computer Vision (ICCV)*, pages 15118–15129. IEEE, 2021. 2

[30] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *International Conference on Computer Vision (ICCV)*, pages 2720–2727. IEEE, 2013. 2, 4

[31] Weixin Luo, Wen Liu, and Shenghua Gao. Remembering history with convolutional lstm for anomaly detection. In *International conference on multimedia and expo (ICME)*, pages 439–444. IEEE, 2017. 2

[32] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *International Conference on Computer Vision (ICCV)*, pages 341–349. IEEE, 2017. 2, 4

[33] Weixin Luo, Wen Liu, and Shenghua Gao. Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection. *Neurocomputing*, 444:332–337, 2021. 2

[34] Osama Makansi, Özgün Cicek, Yassine Marrakchi, and Thomas Brox. On exposing the challenging long tail in future prediction of traffic actors. In *International Conference on Computer Vision (ICCV)*, pages 13147–13157. IEEE, 2021. 2

[35] Amir Markovitz, Gilad Sharir, Itamar Friedman, Lihi Zelnik-Manor, and Shai Avidan. Graph embedded pose clustering for anomaly detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10539–10547. IEEE, 2020. 1, 2, 4, 5

[36] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11996–12004. IEEE, 2019. 1, 2, 3, 4, 5, 7, 8

[37] Rashmiranjan Nayak, Umesh Chandra Pati, and Santos Kumar Das. A comprehensive review on deep learning-based methods for video anomaly detection. *Image and Vision Computing*, 106:104078, 2021. 1

[38] Duc Tam Nguyen, Zhongyu Lou, Michael Klar, and Thomas Brox. Anomaly detection with multiple-hypotheses predictions. In *International Conference on Machine Learning (ICML)*, pages 4800–4809. PMLR, 2019. 2

[39] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11205–11214. IEEE, 2021. 2

[40] Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. Probabilistic representations for video contrastive learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14711–14721. IEEE, 2022. 2

[41] Parvaneh Parvin, Fabio Paternò, and Stefano Chessa. Anomaly detection in the elderly daily behavior. In *International Conference on Intelligent Environments (IE)*, pages 103–106. IEEE, 2018. 1

[42] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding*, 208:103219, 2021. 2

[43] Helei Qiu, Biao Hou, Bo Ren, and Xiaohua Zhang. Spatio-temporal tuples transformer for skeleton-based action recognition. *arXiv preprint arXiv:2201.02849*, 2022. 4

[44] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, and Nicu Sebe. Abnormal event detection in videos using generative adversarial nets. In *International Conference on Image Processing (ICIP)*, pages 1577–1581. IEEE, 2017. 2

[45] Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, and Reinhard Klette. Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Transactions on Image Processing*, 26(4):1992–2004, 2017. 2

[46] Vishal Sharma, Ravinder Kumar, Wen-Huang Cheng, Mohammed Atiquzzaman, Kathiravan Srinivasan, and Albert Y Zomaya. Nhad: Neuro-fuzzy based horizontal anomaly detection in online social networks. *IEEE Transactions on Knowledge and Data Engineering*, 30(11):2171–2184, 2018. 1

[47] Adrian Spurr, Aneesh Dahiya, Xi Wang, Xucong Zhang, and Otmar Hilliges. Self-supervised 3d hand pose estimation from monocular rgb via contrastive learning. In *International Conference on Computer Vision (ICCV)*, pages 11230–11239. IEEE, 2021. 2

[48] Alexandros Stergiou and Ronald Poppe. Analyzing human–human interactions: A survey. *Computer Vision and Image Understanding*, 188:102799, 2019. 2

[49] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6479–6488. IEEE, 2018. 1, 2

[50] Dídac Surís, Ruoshi Liu, and Carl Vondrick. Learning the predictability of the future. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12607–12617. IEEE, 2021. 2

[51] Dídac Surís and Carl Vondrick. Representing spatial trajectories as distributions. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 13731–13744, 2022. 2

[52] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *International Conference on Computer Vision (ICCV)*, pages 4975–4986. IEEE, 2021. 2

[53] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10078–10093. IEEE, 2022. 2

[54] Zhiwei Wang, Zhengzhang Chen, Jingchao Ni, Hui Liu, Haifeng Chen, and Jiliang Tang. Multi-scale one-class recurrent neural networks for discrete event sequence anomaly detection. In *Conference on Knowledge Discovery & Data Mining (SIGKDD)*, pages 3726–3734. ACM, 2021. 2

[55] Sarthak Yadav and Neil Zeghidour. Learning neural audio features without supervision. In *Interspeech*, 2022. 2

[56] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI Conference on Artificial Intelligence*. AAAI, 2018. 2

[57] Di Yang, Yaohui Wang, Quan Kong, Antitza Dantcheva, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Self-supervised video representation learning via latent time navigation. In *AAAI Conference on Artificial Intelligence*. AAAI, 2023. 2

[58] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision (ECCV)*, pages 346–364. Springer, 2020. 2

[59] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. Spatio-temporal autoencoder for video anomaly detection. In *International Conference on Multimedia*, pages 1933–1941. ACM, 2017. 2

[60] Andrea Ziani, Zicong Fan, Muhammed Kocabas, Sammy Christen, and Otmar Hilliges. Tempclr: Reconstructing hands via time-coherent contrastive learning. In *International Conference on 3D Vision (3DV)*, pages 627–636. IEEE, 2022. 2