Holistic Representation Learning for Multitask Trajectory Anomaly Detection Supplementary Material

Table S1. **Encoder hyperparameter optimization on HR-STC**. The setting with the best overall score is in light gray. The best AUC score per task is in **bold**. We note that λ_s (for joints) and λ_{bb} (for box corners) are only used for trajectory reconstruction.

lr	β	γ	λ_s	λ_{bb}	Ftr	Prs	Pst
$5e^{-4}$	0.001	0.1	2	4	75.6	72.8	74.7
	0.001	0.1	3	5	75.9	72.9	75.0
$1e^{-4}$	1.0	0.1	3	5	76.9	73.1	75.1
	0.1	0.1	3	5	77.4	73.2	75.4
	0.01	0.1	3	5	77.7	73.4	75.8
	0.001	1.0	3	5	76.8	72.9	74.5
	0.001	0.1	3	5	**77.9**	73.5	**75.7**
	0.001	0.01	3	5	77.7	**73.6**	75.4

Table S2. **HR-STC AUC scores over different λ hyoperparameters for Ftr**. Best combination of λ_s and λ_{bb} is in green. Settings for which their AUC is above the state-of-the-art are in **bold**. Settings that were not explored are denoted with N/A in gray.

		λ_s							
		1	2	3	4	5	6	7	8
λ_{bb}	1	76.5	N/A	N/A	N/A	76.2	N/A	75.6	75.9
	2	N/A	75.9	77.3	**77.6**	77.3	N/A	76.9	76.7
	3	76.5	76.8	N/A	77.5	77.9	**77.6**	77.2	76.9
	4	N/A	N/A	N/A	76.3	76.5	76.9	77.4	77.2
	5	N/A	N/A	76.4	76.9	76.6	76.4	76.7	76.7
	6	76.1	76.3	76.4	76.6	76.7	76.4	76.2	N/A
	7	76.4	76.2	76.2	76.5	76.8	76.8	77.0	76.9
	8	N/A	N/A	N/A	76.3	N/A	N/A	N/A	N/A

S1. Encoder and decoder hyperparameters

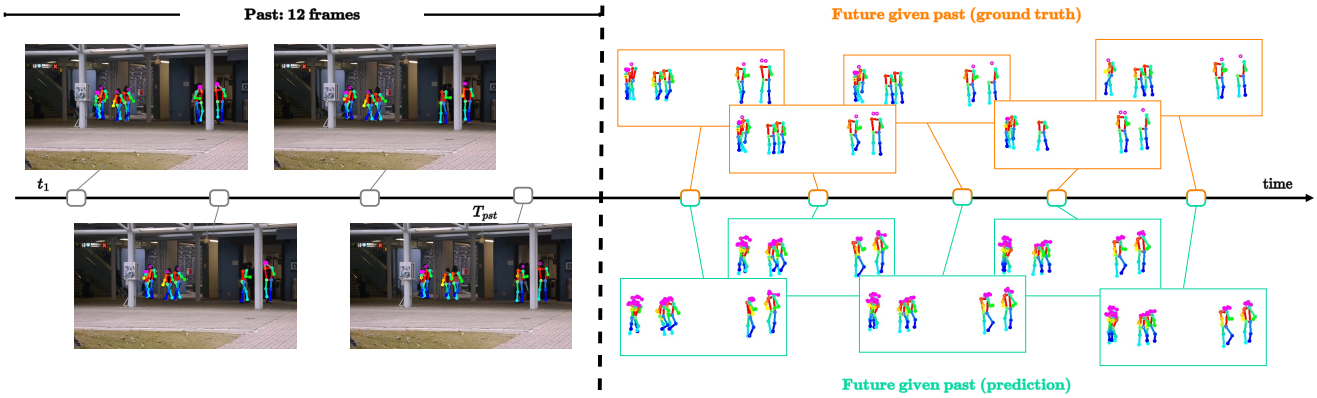
We discover the optimal hyperparameters with random search. For the encoder hyperameters we define a search space of $\beta \in \{1.0, 0.1, 0.01, 0.001\}$ and $\gamma \in \{1.0, 0.1, 0.01\}$ where β is the hyperparameter used by the regularizer for soft negative pairs S^- and γ is the margin hyperparameter used in the triplet loss. A full list of searched combinations is shown in Table S1. The largest decreases in performance are observed for $\beta = 1.0$ corresponding to strong penalization for S^- pairs. As we also motivate in Section 4.2, adjacent points from the same trajectory are bound to include similarities therefore strong penalization pushes representations of adjacent points further apart.

We additionally perform a random search over λ decoder hyperparameters for skeleton joints, denoted as λ_s, and bounding box corners, denoted as λ_{bb}. As shown in Table S2, the general trend for the settings with AUC scores comparable to those of the state-of-the-art is $\lambda_{bb} < \lambda_s$ and $\lambda_{bb} < 5$. We believe that $\lambda_{bb} < \lambda_s$ is due to the trajectories of bounding box corners significantly varying based on the orientation of the pose being tracked. In contrast, the trajectories of spatial locations of skeleton joints are constrained to the corresponding body parts.

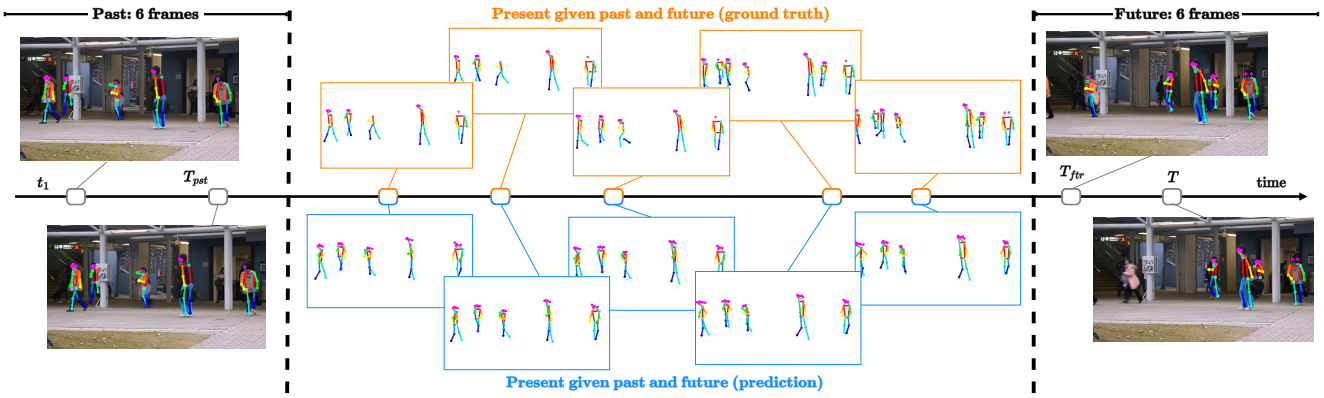
S2. Additional predicted segment examples

Our proposed multitask approach for anomaly detection is based on the holistic representation of trajectories as segments. Crucially, the continuity of trajectories is based on the detection of skeletons at the pose level. Skeletons or individual keypoints not detected by the pose detector would correspond to partial ground truth trajectories for which their reconstructions are not possible. As with the ablation results in Section 4.3 of the main text, we provide qualitative results for Ftr, Prs, and Pst extrapolation and interpolation tasks.

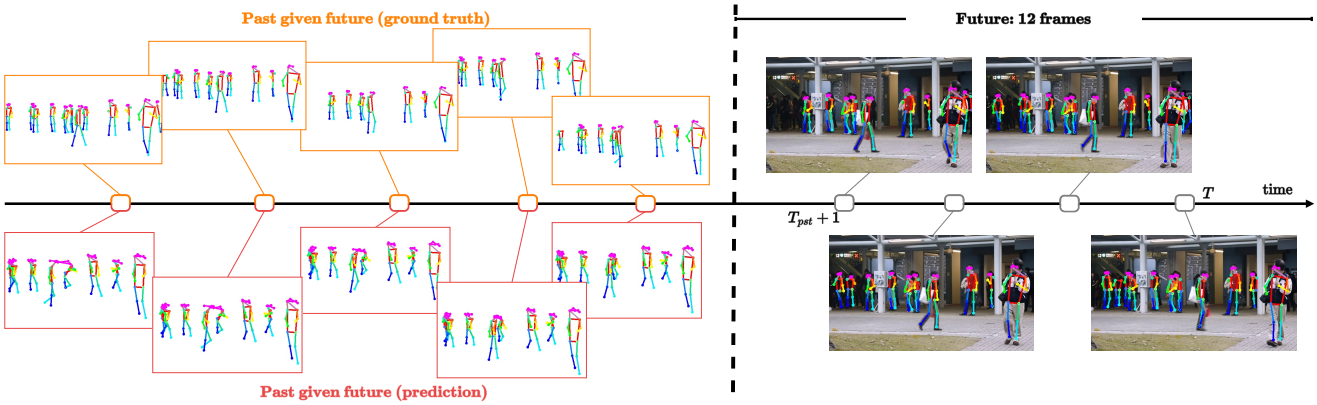
Figure S1 shows examples of non-continuous trajectories and reconstructions. Across all tasks, our model is capable of reconstructing continuous trajectories despite the absence of undetected keypoints in the ground truth. For either extrapolation tasks Ftr and Pst partially observed trajectories are reconstructed for the entire skeletons regardless of the keypoints detected. For cases in which entire skeletons are not detected, reconstructions are also not created. Specifically, the leftmost skeletons in the Pst example are not detected in the majority of the observable frames and in turn, cannot be reconstructed/predicted. In contrast, predictions for absent keypoints, as shown in the Ftr example, can still be made as the model learns expected motions with respect to the remaining observable keypoints in the trajectory. In example of the Prs interpolation task, reconstructions are only done for skeletons that are correctly detected in both proceeding or succeeding frames. As shown, predictions given learned motions are also made for the joints that are not observable.



(a) **Predicting the future given the past** (Ftr).



(b) **Predicting the present given both the past and future** (Prs).



(c) **Predicting the past given the future** (Pst).

Figure S1. **Skeleton sequence reconstruction in HR-Avenue**. Input sequences are keypoints from 18 frames with each of the predicted segments for Ftr, Prs, and Pst task being 6 frames. In each of the occluded segments, not all keypoints are recognized by the pose detector. This relates to partial ground truths. As shown the model can learn strong representations with predicted keypoints still being sensible despite not being detected and part of the ground truth.

S3. Ablations on reconstruction lengths

Motivated by Table 3 in the main paper, we further jointly ablate over occluded segment lengths for training and inference in Table S3. Normally, the model is optimized to extrapolate/interpolate occluded segments of predefined length. However, scenarios may exist in which retraining or finetuning for specific lengths may not be available. We thus train our model with occluded segment lengths of either $|\hat{\mathbf{t}}| = 6$ or $|\hat{\mathbf{t}}| = 12$. In each setting, the encoder learns to pull closer latent and learned representations for the occluded segments while the decoder learns to reconstruct the full trajectory of $|\mathbf{t}| = 18$. For each setting, we run infer-

Table S3. **AUC scores on HR-STC and HR-Avenue over varying occluded segment lengths for training and inference**. The default settings in which the occluded segment lengths are the same for both training and inference are denoted in gray.

$\uparrow \lvert\widehat{\mathbf{t}}\rvert$	$\downarrow \lvert\widehat{\mathbf{t}}\rvert$	HR-STC			HR-Avenue		
		Ftr	Prs	Pst	Ftr	Prs	Pst
6	6	77.9	73.5	75.7	89.4	86.3	87.6
	12	72.3	69.6	69.8	85.7	83.2	84.4
12	6	72.3	70.6	72.1	85.0	81.3	84.7
	12	75.9	72.7	75.9	87.9	85.5	86.4

ence with either $|\widehat{\mathbf{t}}| = 6$ or $|\widehat{\mathbf{t}}| = 12$ occluded segments. For clarity, training and inference lengths are denoted with $\uparrow |\widehat{\mathbf{t}}|$ and $\downarrow |\widehat{\mathbf{t}}|$ respectively. On average, a -3.1 and -2.6 decrease in the AUC score is observed over both HR-STC and HR-Avenue on the $\uparrow |\widehat{\mathbf{t}}| = 12$ setting when changing the inference occlusion from $\downarrow |\widehat{\mathbf{t}}| = 12$ to $\downarrow |\widehat{\mathbf{t}}| = 6$. A similar drop is observed in the $\uparrow |\widehat{\mathbf{t}}| = 6$ setting when changing $\downarrow |\widehat{\mathbf{t}}| = 6$ to $\downarrow |\widehat{\mathbf{t}}| = 12$ with -5.1 and -3.3 AUC score reductions for HR-STC and HR-Avenue. This demonstrates that despite not being optimized during training, sensible results can still be obtained in inference cases where occluded segments are of varying lengths and finetuning is not possible.