PoseCVAE: Anomalous Human Activity Detection

Yashswi Jain[1], Ashvini K. Sharma[1], V. Rajbabu, Biplab Banerjee

IIT Bombay

yashswijain.kbi@gmail.com

December 8, 2020

[1]Denotes equal contribution

Datasets

- There are three publicly available datasets:

CUHK Avenue[1]	**ShanghaiTech**[2]	**IITB-Corridor**[3]
16 training videos	330 training videos	208 training videos
21 test videos	107 test videos	150 test videos
Short-term anomaly	Multi-camera	Close to real world

Table: Available Datasets

- Since we are concerned only with human anomaly, we will be working on subsets of above datasets i.e., HR-Avenue, HR-ShanghaiTech and HR-IITB Corridor (**proposed by us**)

Preprocessing Pipeline: Trajectory Extraction

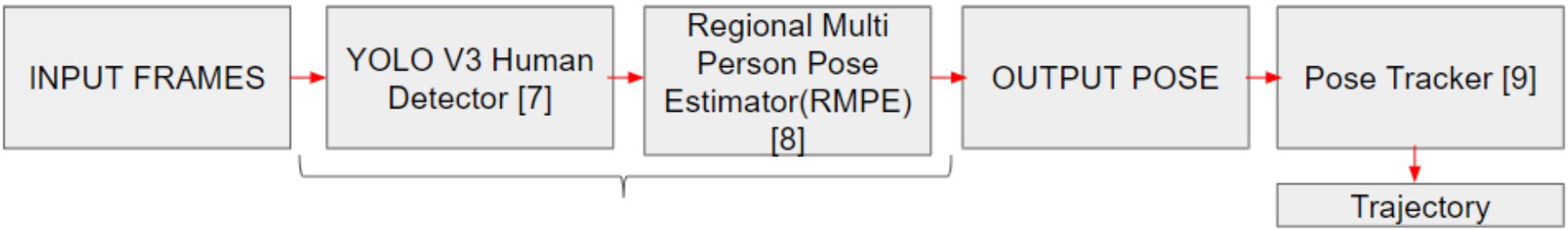


Figure: Trajectory Extraction Pipeline

- Used a network with combined detection and pose estimator
- Pose detector output is 17 keypoints i.e. (x,y) coordinates

Preprocessing Pipeline: Normalization

- Keypoints obtained from pose estimator are not normalized
- This causes increase in error due to closer entities
- To correct depth effect we propose bounding box normalization as shown:

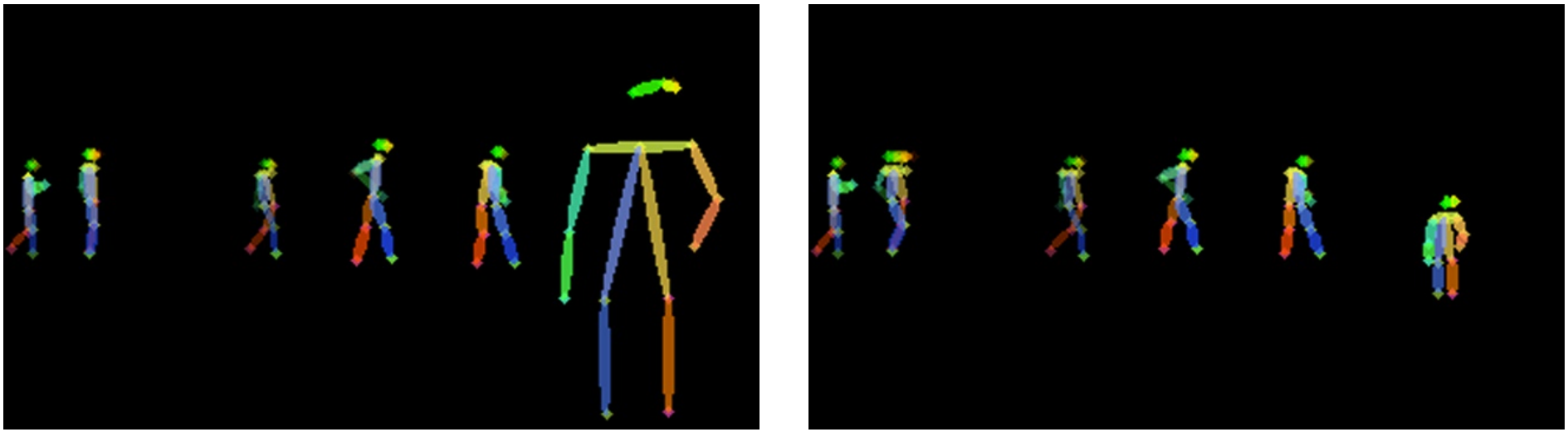


Figure: Left: Without Normalization Right: With Normalization

Preprocessing Pipeline: Data Augmentation

- It's better to take small tracks of person instead of long corrupted tracks
  - The output of Pose tracker is affected by illumination, depth, occlusion
  - Tracking in multi person environment is difficult due to overlapping trajectories
- We use sliding window to divide trajectories into multiple tracks
- This helps in data augmentation as well

PoseCVAE: High Level Flow

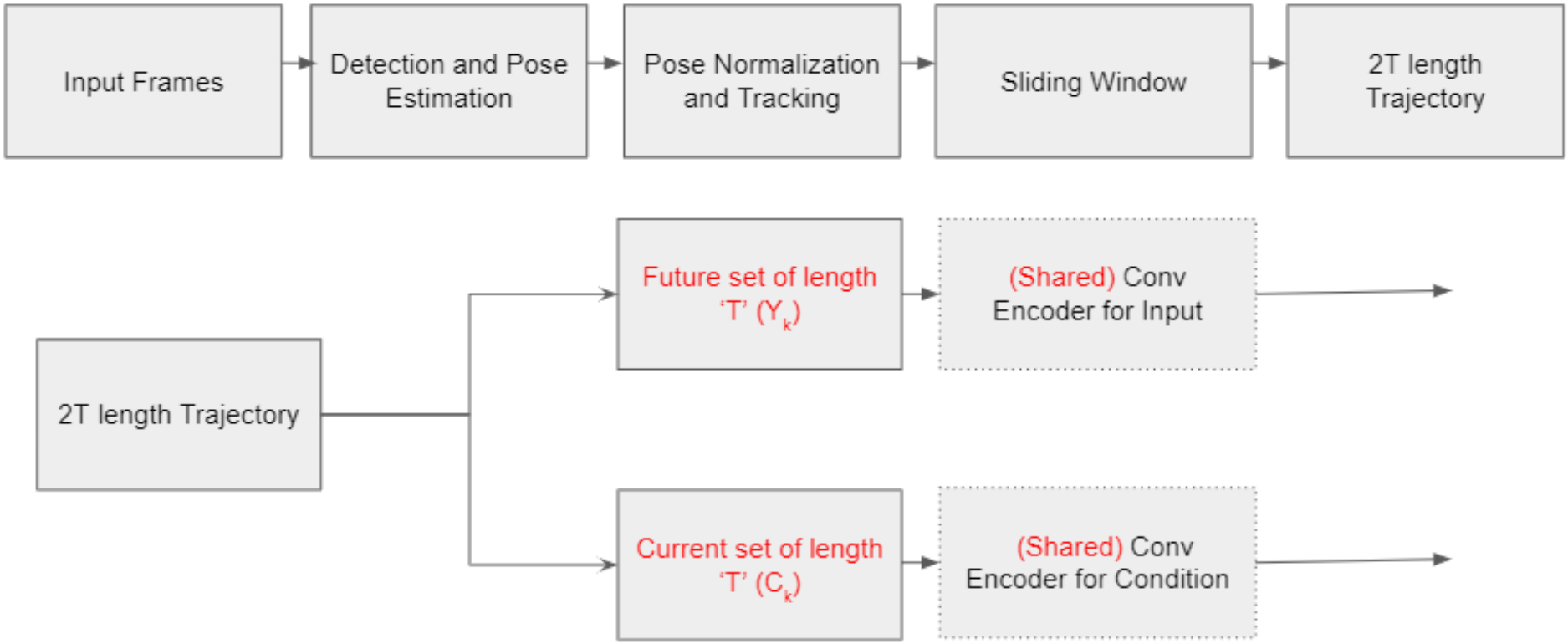


Figure: High Level Flow Diagram

PoseCVAE: Architecture

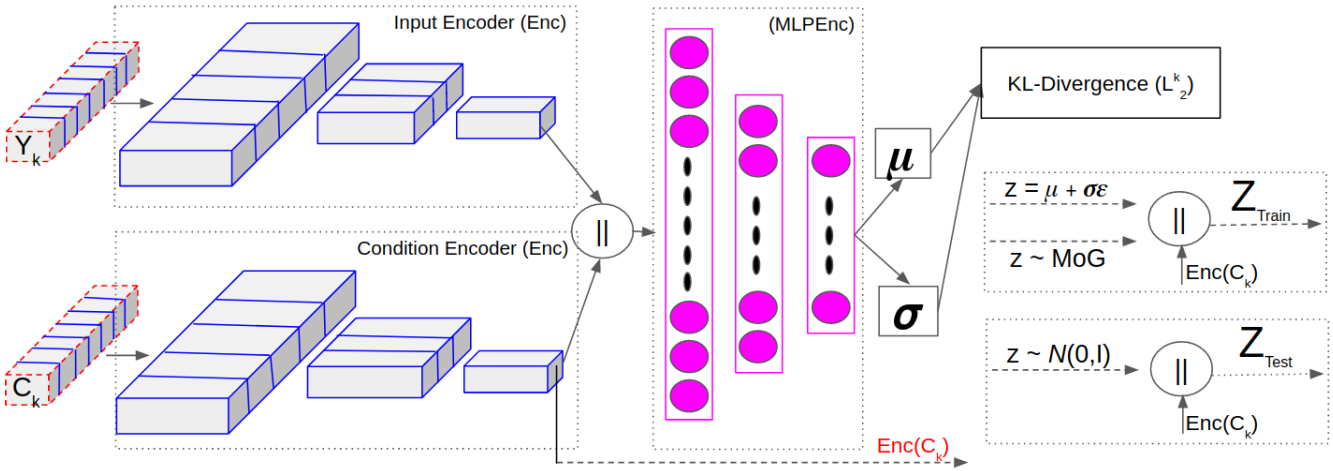


Figure: Encoder

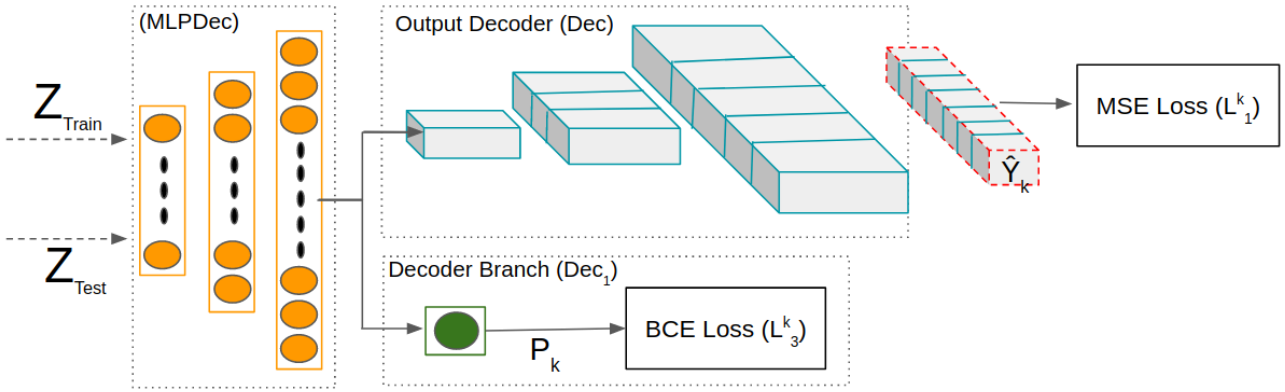


Figure: Decoder

Imitating Abnormal Pose Trajectory in the Latent Space

- To maximise the separation between normal and abnormal classes, we split a decoder branch (Dec_1) which gives class probability, P_k as output
- Normal Class is labelled '0', Abnormal class is labelled '1'
- Different possibilities for the concatenated latent vector:

$$Z_{normal} \equiv z \sim \mathcal{Q}(.) \mid\mid Enc(C_k) \tag{1}$$

$$Z_{abnormal} \equiv z \sim \mathcal{N}(0, I) \mid\mid Enc(C'_k) \tag{2}$$

$$\tilde{Z}_{abnormal} \equiv z \sim MoG \mid\mid Enc(C_k) \tag{3}$$

- The output of the classifier branch is mapped as follows:

$$Dec_1(MLPDec(Z_{normal})) \rightarrow 0 \tag{4}$$

$$Dec_1(MLPDec(\tilde{Z}_{abnormal})) \rightarrow 1 \tag{5}$$

Loss Function

Used combination of three loss functions during training:

- Reconstruction Loss: Maximising the conditional expectation translates into minimising MSE:

$$L_1^k(Y_k, \hat{Y}_k) = \left|\left|\hat{Y}_k - Y_k\right|\right|_2^2 \tag{6}$$

- KL divergence Loss: Minimise the KLD to maximise the conditional likelihood:

$$L_2^k(\mu, \sigma) = \mathcal{KL}[\mathcal{N}(\mu(Y_k, C_k), \sigma(Y_k, C_k)) \,||\, \mathcal{N}(0, I)] \tag{7}$$

- BCE loss: **To make normal and abnormal latent samples more distinguishable:**

$$L_3^k(y_k, P_k) = -(y_k \log P_k + (1 - y_k) \log(1 - P_k)) \tag{8}$$

Training Strategy

- Input: future trajectory to be predicted, length = 'T'
- Condition: past trajectory of length 'T'
- Aim: learn conditional posterior and reconstruct the input given the condition

We train in 3 stages:

- Stage 1: Self Supervised Learning (Pre-training the Conv. Encoder and decoder)
  - Objective: Reconstruct the given trajectories
- Stage 2: Unsupervised Learning (Training the PoseCVAE)
  - Objective: Reconstruct the given trajectory given the past trajectory and minimise the KLD (Maximising the conditional likelihood)
- **Stage 3: Unsupervised with OoD sample generation and minimise BCE** (Fine-tuning the PoseCVAE framework)
  - Objective: For normal latent points: Minimise the KLD, MSE and BCE, for abnormal latent points: Minimise the MSE and BCE

Inference: Obtaining the Frame-level Anomaly Score

- Input: Noise randomly sampled from standard normal
- Condition: past trajectory, length = 'T'
- Output: future trajectory, length = 'T'
- Obtain the corresponding squared difference between prediction and GT
- Average it to obtain the final squared difference for a given time instant $(T+1)$ and a given person (k), $\delta_k(T+1)$
- Obtain $\delta_k(i) \forall\, i \,\epsilon\, T_k, \forall k$. Here T_k is the entire track of person 'k'
- Frame-level anomaly score, $\Delta(t_0)$, at $t = t_0$, is obtained as shown:

$$\Delta(t_0) = \max_{j \in \mathcal{S}(t_0)} \delta_j(t_0) \tag{9}$$

Here $S(t_0)$ refers to the set of all person IDs that appear in the video at $t = t_0$

Results: Visualisation

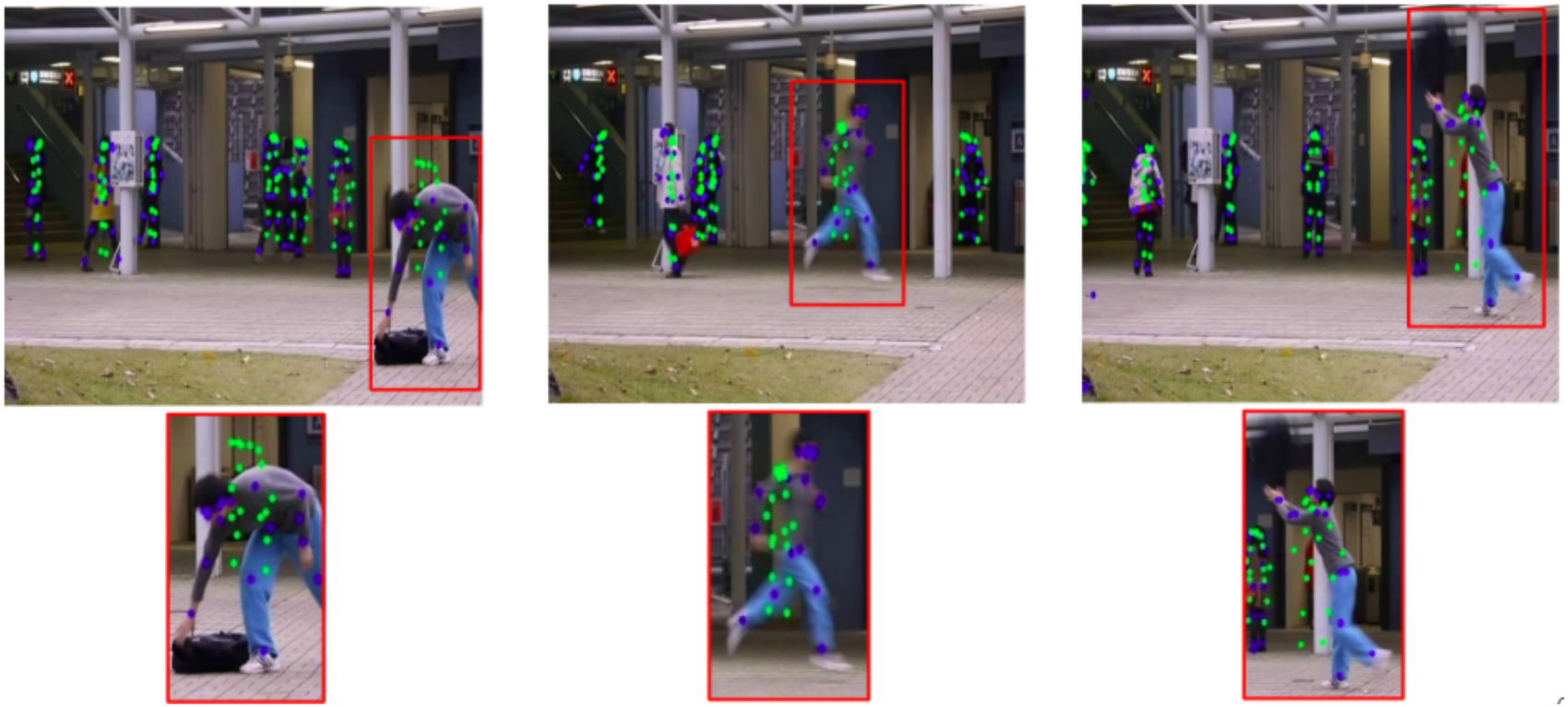


Figure: Green skeleton is from the predicted trajectory and Blue one is from the ground truth. Notice the greater dissimilarity between the two skeletons for abnormal motion/ poses.

Results: Frame-level Anomaly Score Plot

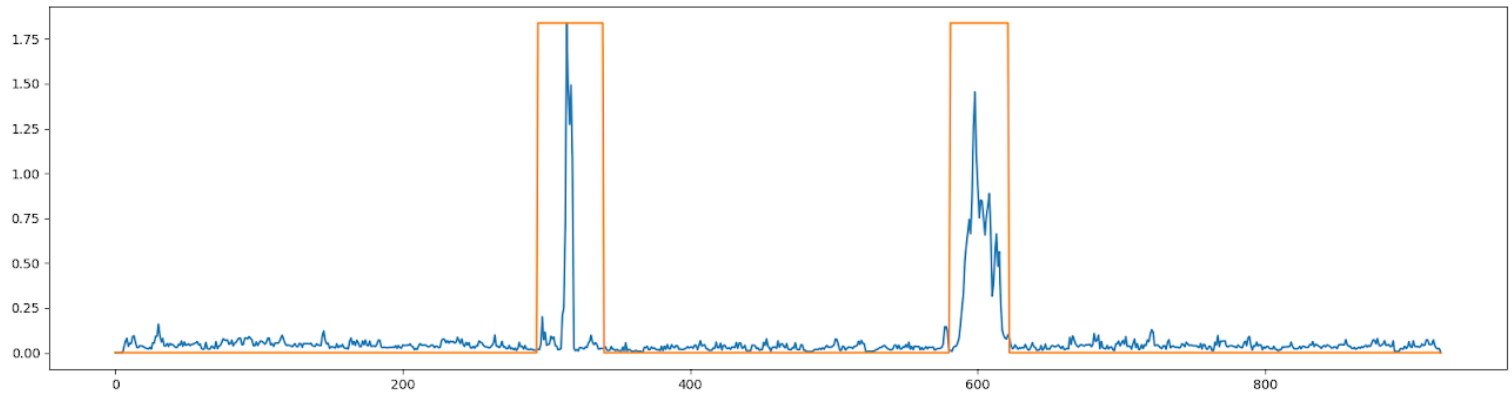


Figure: Frame-level anomaly score plot, Video 3 from the test set, HR-Avenue (HR version of Avenue Dataset[1]). Notice the frame-level anomaly score is lower for normal frames and shoots up for abnormal frames.

Results: AUC Score

	HR-Avenue	HR-ShanghaiTech	HR-IITB
Hasan *et al.* [4]	84.80	69.80	-
Liu *et al.* [2]	86.20	72.70	-
Luo *et al.* [5]	-	-	-
Morais *et al.* [6]	86.30	75.40	68.07
Rodrigues *et al.* [3]*	88.33	**77.04**	-
Ours	**87.78**	75.86	**70.60**

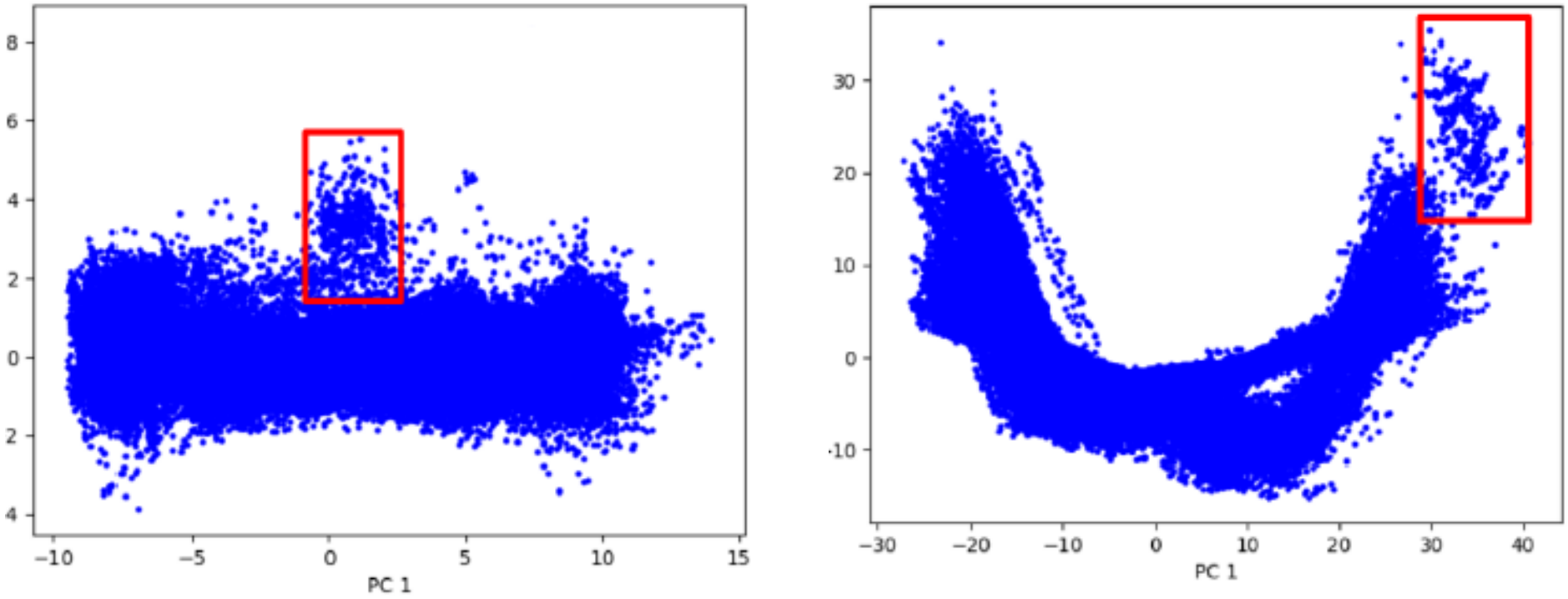
	Avenue	ShanghaiTech	IITB Corridor
Hasan *et al.* [4]	70.20	69.80	-
Liu *et al.* [2]	**84.90**	72.80	64.65
Luo *et al.* [5]	81.71	-	68.00
Morais *et al.* [6]	-	73.40	64.27
Rodrigues *et al.* [3]*	82.85	**76.03**	67.12
Ours	82.10	74.90	**67.43**

Results: Comparison with the SOTA

	T	Predictions/Iteration	AUC
Rodrigues *et al.*[3] **(Multi-timescale)**	3 only	2	72.05%
	3 & 5	4	73.39%
	3,5 & 13	6	75.65%
	3,5,13 & 25	8	77.04%
Ours **(One-timescale)**	7	**1**	75.86%

Table: Frame-level AUC score comparison between [3] and Our method on HR-ShanghaiTech for different timescales

Ablation Study: Effect of Multistage Training



(a) Trained without Stage 3

(b) Trained with Stage 3

Figure: Latent space representation of the test set trajectories obtained from PoseCVAE post- training completion. Notice the increase in the separation between the normal and abnormal trajectory classes after introduction of stage 3 in the training strategy.

Shortcomings: Videos in which we do not perform well

- We do not perform well on Video 25 and Video 32 of HR-Avenue
- We achieve AUC scores of $\approx 36\%$ and $\approx 44\%$ respectively on both of the videos
- If we calculate the AUC score excluding the two videos, we get an overall AUC score of 90.87% (Current: 87.78%)

Reason: Both these videos, the anomaly is due to the person roaming in a restricted area.

- **We accept this as a shortcoming of our framework as we capture only pose based information and discard locality during pre-processing**

HR-IITB Corridor Dataset

We remove all the object driven anomalies from the original IITB-Corridor dataset proposed in [3]

- Test split: 150 videos, Train split: 208 videos
- Activities grouped into different categories such as: Fighting, Chasing, Cycling, Loitering, Sudden Running, Protests, Carrying Objects, Bag Exchange, Playing with a ball and Hiding
- We construct the HR-IITB dataset as shown:
  - Included Videos from IITB-Corridor: Fighting, Chasing, Cycling, Loitering, Sudden Running and Protests. A total of 78 videos
  - Removed Videos from IITB-Corridor: Carrying Objects, Bag Exchange, Playing with a ball and Hiding. A total of 72 videos

References

Lu, Cewu and Shi, Jianping and Jia, Jiaya (ICCV 2013)

Abnormal Event Detection at 150 FPS in MATLAB

Liu, Wen and Luo, Weixin and Lian, Dongze and Gao, Shenghua (CVPR 2018)

Future frame prediction for anomaly detection - a new baseline

Rodrigues, Royston and Bhargava, Neha and Velmurugan, Rajbabu and Chaudhuri, Subhasis (WACV 2020)

Multi-timescale Trajectory Prediction for Abnormal Human Activity Detection

Hasan, Mahmudul and Choi, Jonghyun and Neumann, Jan and Roy-Chowdhury, Amit K and Davis, Larry S (CVPR 2016)

Learning temporal regularity in video sequences

Luo, Weixin and Liu, Wen and Gao, Shenghua (ICCV 2017)

A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework

Morais, Romero and Le, Vuong and Tran, Truyen and Saha, Budhaditya and Mansour, Moussa and Venkatesh, Svetha (CVPR 2019)

Learning Regularity in Skeleton Trajectories for Anomaly Detection in Videos

References

Joseph Redmon and Santosh Kumar Divvala and Ross B. Girshick and Ali Farhadi (CVPR 2016)

You Only Look Once: Unified, Real-Time Object Detection

Haoshu Fang and Shuqin Xie and Cewu Lu (ICCV 2017)

RMPE: Regional multi-person pose estimation

Yuliang Xiu and Jiefeng Li and Haoyu Wang and Yinghong Fang and Cewu Lu (BMVC 2018)

Poseflow: Efficient online pose tracking

The End