# PoseCVAE: Anomalous Human Activity Detection

Yashswi Jain*, Ashvini Kumar Sharma*, Rajbabu Velmurugan and Biplab Banerjee
Indian Institute of Technology Bombay, India 400076

## Preprocessing Pipeline: Trajectory Extraction

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

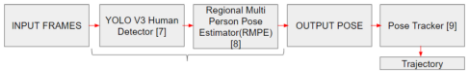



Figure: Trajectory Extraction Pipeline

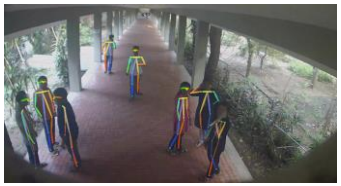


Figure: Output pose from RMPE

## 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:

$$Y_{out} = \left(\frac{Y_{pred} - Y_c}{Width}\right) Y_{hyp} + Y_{centre}$$

$$X_{out} = \left(\frac{X_{pred} - X_c}{Height}\right) X_{hyp} + X_{centre}$$

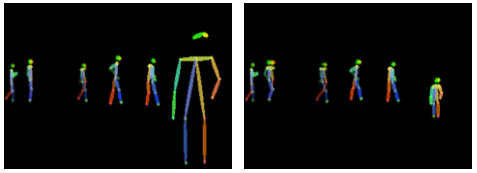
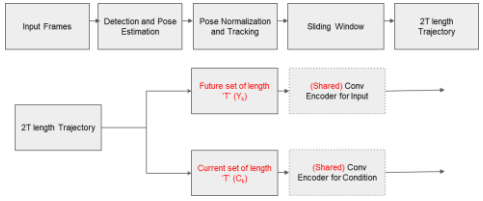


Figure: Left: Without Normalization Right: With Normalization

## High Level Flow Diagram




## PoseCVAE: Architecture

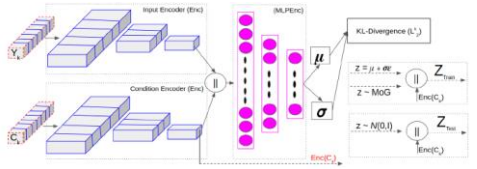



Figure: Encoder

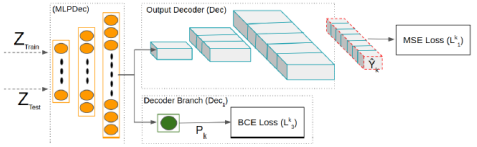



Figure: Decoder

## Imitating Abnormal Pose in 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:

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

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

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

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

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

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

## Loss Function

Used combination of three loss functions during training:

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

$$L^k_1(Y_k, \hat{Y}_k) = \left|\left|\hat{Y}_k - Y_k\right|\right|^2_2 \quad (6)$$

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

$$L^k_2(\mu, \sigma) = \mathcal{KL}[\mathcal{N}(\mu(Y_k, C_k), \sigma(Y_k, C_k)) \mid\mid \mathcal{N}(0, I)] \quad (7)$$

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

$$L^k_3(y_k, P_k) = -(y_k \log P_k + (1 - y_k) \log(1 - P_k)) \quad (8)$$

## 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: 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:

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

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

## Result: 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** |

| | $T$ | 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

## Results: Frame Level Anomaly Score

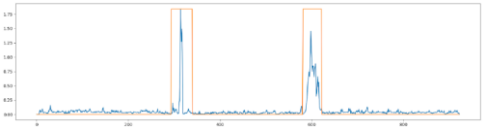
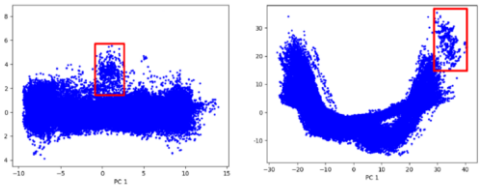


Figure: Frame-level anomaly score plot, Video 3 from the test set, HR-Avenue (HR version of Avenue Dataset[1]). Notice the frame-level anomaly score is lower for normal frames and shoots up for abnormal frames.

## Result: Visualization

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.

## 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.

## References

[1] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab," in Proceedings of the IEEE international conference on computer vision, pp. 2720–2727, 2013.
[2] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection–a new baseline," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6536–6545, 2018.
[3] R. Rodrigues, N. Bhargava, R. Velmurugan, and S. Chaudhuri, "Multitimescale trajectory prediction for abnormal human activity detection," in Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 2626–2634, 2020
[4] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 733–742, 2016.
[5] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked rnn framework," in The IEEE International Conference on Computer Vision (ICCV), Oct 2017.
[6] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh, "Learning regularity in skeleton trajectories for anomaly detection in videos," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11996–12004, 2019.
[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788, 2016.
[8] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in Proceedings of the IEEE International Conference on Computer Vision, pp. 2334–2343, 2017.
[9] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, "Pose flow: Efficient online pose tracking," arXiv preprint arXiv:1802.00977, 2018.

* These authors contributed equally