Normalizing Flows for Human Pose Anomaly Detection Supplementary Material

Or Hirschorn
Tel-Aviv University, Israel
orhirschorn@mail.tau.ac.il

Shai Avidan
Tel-Aviv University, Israel
avidan@eng.tau.ac.il

1. Introduction

The following sections include additional information about our method.

Section 2 presents further ablation studies conducted to evaluate our model. Section 3 provides implementation details of our method, and Section 4 describes the implementation details of the baseline methods used. Section 5 provides examples of our method's performance from the ShanghaiTech and UBnormal datasets.

2. Ablation Study - Cont.

In this section, we provide further ablation experiments used to evaluate different model components:

Partial Data Training: Acquiring normal training data might not be easy. Thus, evaluating the performance of our method using limited training data is essential. We use different sized subsets of the ShanghaiTech dataset S_i, where $S_i \subset S_{i+1}$ for $|S_i| < |S_{i+1}|$. Figure 1 shows our method's performance using the different subsets. Our model losses less than 3% AUC using only 10% of the training data. This is a side benefit of our compact model, which includes only $\sim 1K$ parameters. In addition, as most actions considered normal in the dataset are walking, a small portion of the dataset is enough to assign high probability to this aciton. Thus, limited training data suffices for near state-of-the-art performance.

Segment Window: We explore the effect of different pose segment windows τ on the performance of our model. As can be seen in Figure 2, using longer segment windows results in better performance, up until a time limit $\tau = T$. We believe this limit is affected by the performance of the human pose estimator and tracker. In general, a larger τ is desirable, so complex actions could be better learned. Thus, we conclude that using a larger segment window is preferable, subject to the pose estimation and tracking ability to output long human pose segments.

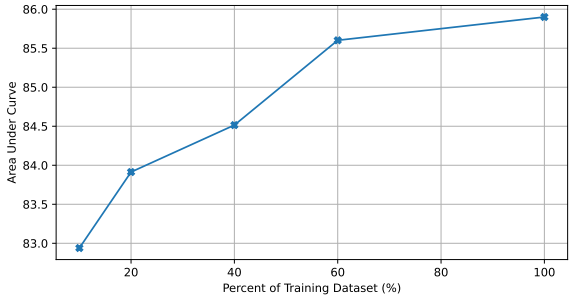


Figure 1. **AUC Score for Partial Data Training:** Performance of unsupervised models trained on ShanghaiTech for different percentages of training data. Our model achieves near state-of-the-art performance using only a limited amount of training samples.

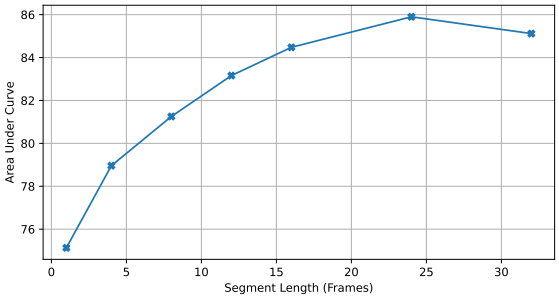


Figure 2. **AUC Score for Different Segment Lengths:** Performance of unsupervised models trained on ShanghaiTech with different segment lengths. A larger segment window is preferable, subject to the pose estimation and tracking ability to output long human pose segments.

Number of Flow Layers: We explore our model's performance using different flow layers K on the ShanghaiTech dataset. As demonstrated in Figure 3, using only $K = 8$ results in state-of-the-art performance. We believe that for more complex datasets, a larger K might be needed.

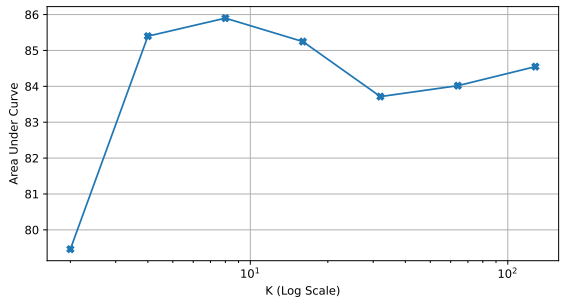


Figure 3. **AUC Score for Different Numbers of Flows:** Performance of unsupervised models trained on ShanghaiTech with a different number of flows K. $K = 8$ results in state-of-the-art performance.

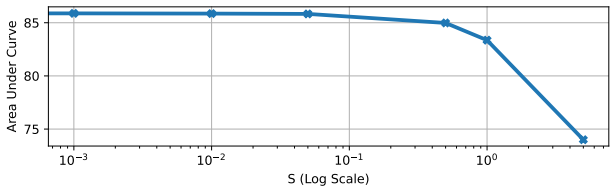


Figure 4. **AUC for Noisy Keypoints:** Performance of models trained on ShanghaiTech, adding various scales of Gaussian noise.

Pose Estimation Errors: To analyze the impact of the pose estimators' errors, we modeled the estimators errors by adding to each keypoint Gaussian noise. After the pose normalization, we added varying scales S of noise $S \cdot z$ where $z \sim \mathcal{N}(0, I)$. As shown in Figure 4, our model is robust to a significant amount of keypoints noise, which implies good performance when using less accurate pose estimators.

3. Implementation Details

Our method is implemented in PyTorch, and all experiments were conducted on a single NVIDIA Titan Xp GPU.

Architecture. We observed the best results when using the unit adjacency matrix in the affine layers, where each joint has an equal effect on the others. We use a segment window $\tau = 24$ frames for ShanghaiTech and $\tau = 16$ frames for UBnormal, as the results of the pose estimation and tracking of UBnormal are less accurate than ShaghaiTech.

Training. We optimize the normalizing flows network parameters with an Adam optimizer, learning rate $5 * 10^{-4}$, momentum 0.99, for eight epochs with batch size 256.

For the unsupervised setting, we use the prior $\mathcal{N}(3, I)$, and for the supervised setting we use $\mathcal{N}(10, I)$ for normal samples and $\mathcal{N}(-10, I)$ for abnormal samples, ensuring:

$$|\mu_{normal} - \mu_{abnormal}| \gg 0$$

4. Baseline Implementation Details

The evaluation of the pose-based methods was conducted using their publicly available implementation[1][2][3][4][5]. The training was done using the same pose data.

Similarly, the evaluation of the Jigsaw anomaly detection was conducted using their implementation[6]. The training was done using default parameters used by the authors, and changes were only made to adapt the data loading portion of the models to our datasets.

As we couldn't reproduce some of the results, thus we took the original AUC scores from the original paper. In addition, some results were taken from [1].

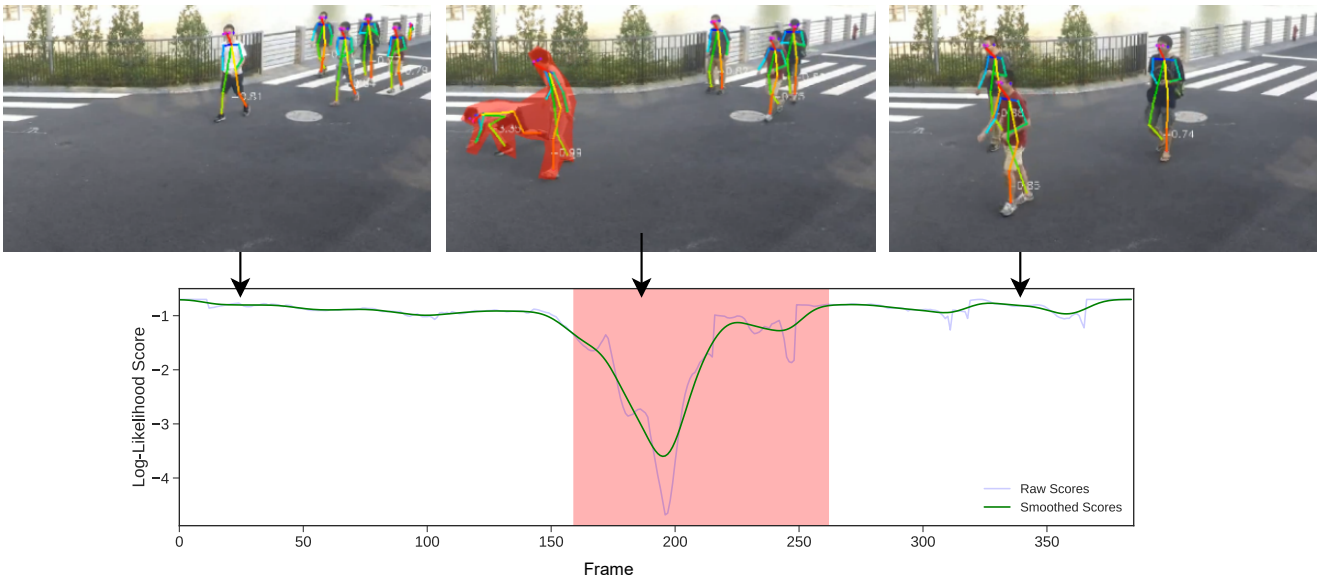
5. Additional Results

In this section, we showcase more frame-level AUC performance of our model. Figure 5 shows violent behaviors in the ShanghaiTech datasets. Figure 6 shows people passing out or dancing (considered anomalous behavior) from the UBnormal dataset; our method successfully recognizes the anomalies despite the difficult viewpoint.

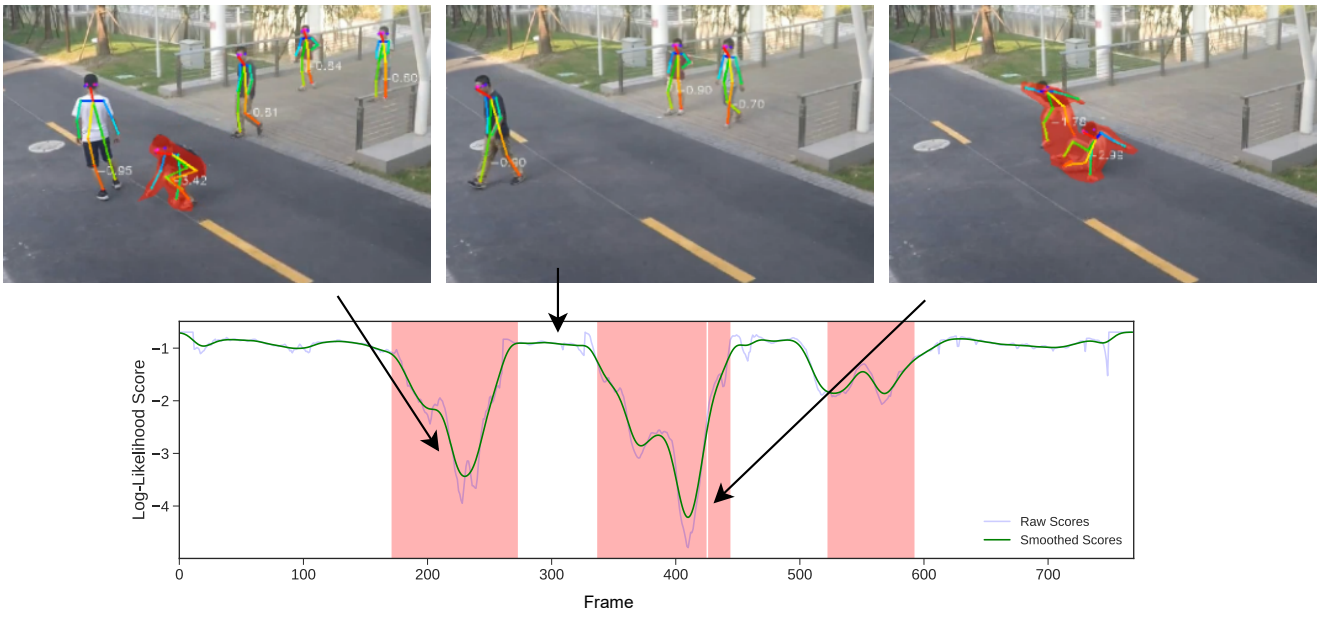
References

[1] Antonio Barbalau, Radu Tudor Ionescu, Mariana-Iuliana Georgescu, Jacob Dueholm, Bharathkumar Ramachandra, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B. Moeslund, and Mubarak Shah. Ssmtl++: Revisiting self-supervised multi-task learning for video anomaly detection, 2022. 2

[1] https://github.com/amirmk89/gepc
[2] https://github.com/yysijie/st-gcn
[3] https://github.com/kchengiva/Shift-GCN
[4] https://github.com/lshiwjx/2s-AGCN
[5] https://github.com/kenziyuliu/ms-g3d
[6] https://github.com/gdwang08/Jigsaw-VAD

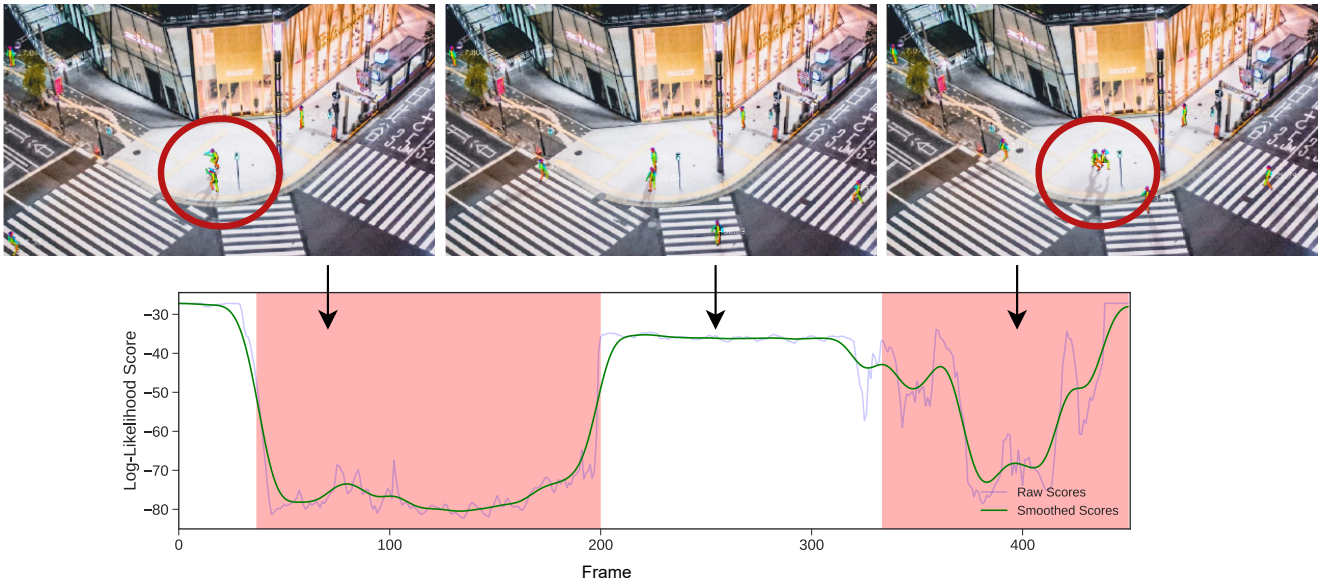


(a)



(b)

Figure 5. **Score examples for ShanghaTech dataset video**. Ground truth anomalous frames and people are marked red. Our unsupervised method is able to correctly recognize strange behavior and violence in both time and space.



(a)

(b)

Figure 6. **Score examples for UBnormal dataset video**. Ground truth anomalous frames are marked red, and anomalies in photos are marked in a red circle. (a) A scene of people dancing - an abnormal behavior in the UBnormal dataset. (b) A scene of a person falling to the ground. Our supervised method is able to recognize anomalies even in difficult viewpoints correctly.