Sequential keypoint density estimator: an overlooked baseline of skeleton-based video anomaly detection

Supplementary Material

A. Derivation of the SeeKer optimization objective

We train the SeeKer model by minimizing the sequence likelihood (6) that boils down to minimizing the Mahalanobis distance between the observed keypoint $X_{t,n}$ and the multivariate normal distribution with parameters $\boldsymbol{\mu}_\theta$ and Σ_θ, alongside a regularization term. The step-by-step derivation goes as follows:

$$L(\theta;\mathcal{D}) = -\sum_{\mathbf{X}\in\mathcal{D}} \ln p_\theta(\mathbf{X}) \tag{10}$$

$$= -\sum_{\mathbf{X}\in\mathcal{D}}\sum_{t=1}^{T}\sum_{n=1}^{N} \ln \mathcal{N}(X_{t,n}|\boldsymbol{\mu}_\theta,\Sigma_\theta) \tag{11}$$

$$= -\sum_{\mathbf{X},t,n} \ln \frac{1}{(2\pi)|\Sigma_\theta|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{X}_{t,n}-\boldsymbol{\mu}_\theta)^T\Sigma_\theta^{-1}(\mathbf{X}_{t,n}-\boldsymbol{\mu}_\theta)\right] \tag{12}$$

$$= \frac{1}{2}\sum_{\mathbf{X},t,n}\left[(\mathbf{X}_{t,n}-\boldsymbol{\mu}_\theta)^T\Sigma_\theta^{-1}(\mathbf{X}_{t,n}-\boldsymbol{\mu}_\theta)\right] + \ln(2\pi) + \ln\det\Sigma_\theta \tag{13}$$

$$\cong \sum_{\mathbf{X},t,n}\left[(\mathbf{X}_{t,n}-\boldsymbol{\mu}_\theta)^T\Sigma_\theta^{-1}(\mathbf{X}_{t,n}-\boldsymbol{\mu}_\theta)\right] + \ln\det\Sigma_\theta \tag{14}$$

In practice, our main model used diagonal covariance, $\Sigma = \boldsymbol{\sigma} I$ which yields:

$$L(\theta;\mathcal{D}) = \sum_{\mathbf{X},t,n}\left[(\mathbf{X}_{t,n}-\boldsymbol{\mu}_\theta)^T\text{diag}(1/\boldsymbol{\sigma})(\mathbf{X}_{t,n}-\boldsymbol{\mu}_\theta)\right] + \ln\sigma_x + \ln\sigma_y \tag{15}$$

B. Architectures of autoreregressive density estimators

B.1. Causal fully connected model

Causality constraints can be effectively incorporated into deep fully connected models by strategically masking weight matrices [4, 21]. Given an input vector $\mathbf{x}$ of size $T \times N \times D$ (e.g. a flattened matrix $\mathbf{X}$ containing a skeleton sequence), we define the causal fully connected layer (C-FC$_l$) as:

$$\text{C-FC}_l(\mathbf{x}) = (\mathbf{W}\odot\mathbf{M})\cdot\mathbf{x} + \mathbf{b}. \tag{16}$$

The mask $\mathbf{M}$ enforces causality by setting weights that correspond to current and subsequent input dimensions to zero. By stacking multiple C-FC$_l$ layers, we can construct a deep causal fully connected model (C-FC). Assuming L hidden layers, we can express the C-FC with the following equations:

$$\mathbf{h}_0 = \mathbf{x} \tag{17}$$

$$\mathbf{h}_l = g((\mathbf{W}_l\odot\mathbf{M}_l)\mathbf{h}_{l-1} + \mathbf{b}_l),\ l = \{1,\dots L\} \tag{18}$$

$$\boldsymbol{\mu},\boldsymbol{\sigma} = (\mathbf{W}_o\odot\mathbf{M}_o)\mathbf{h}_L + \mathbf{b}_o \tag{19}$$

Here, g is some non-linear activation function, such as ReLU. We mask the weights of intermediate layers l with blockwise lower triangular matrices $\mathbf{M}^l$:

$$\mathbf{M}^l_{i,j} = \begin{cases} 1, & \text{if } \lfloor j/D\rfloor \le \lfloor i/D\rfloor \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

ensuring that the elements of each keypoint are still codependent. The mask of the output layer $\mathbf{M}^o$ is an upper diagonal

$$\mathbf{M}^o_{i,j} = \begin{cases} 1, & \text{if } \lfloor j/D\rfloor > \lfloor i/D\rfloor \\ 0, & \text{otherwise,} \end{cases} \tag{21}$$

to ensure independent predictions, e.g. computation without direct influence from its own current and all succeeding representation. We duplicate the mask in the output layer since we output two values for each input, e.g. $\boldsymbol{\mu}$ and Σ. For the main experiments that are on the keypoint granularity level, we use $D = 2$, and for the experiments on the skeleton granularity level $D = N \cdot T = 36$ (Table 5). Finally, to ensure that all keypoints contribute equally, the hidden layer dimensions must be a multiple of the input dimension. This guarantees a balanced flow of information across all keypoints during processing.

Figure 8 shows a minimal example of a causal fully-connected model. For example, the predicted parameters of distribution under which the keypoint K_2 (red) should be probable depends on the keypoint K_1 (blue) as highlighted with red weights (see Figure 3 in the main text). To maintain clarity, keypoints of preceding skeletons and bias terms $\mathbf{b}_l$ are omitted from the visualization.

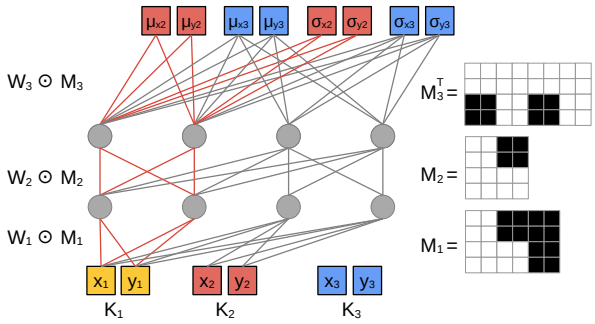

Figure 8. A minimal example of the causal deep fully connected model. Here we causally predict parameters of the distribution for target keypoint K_2 with respect to the preceding keypoint K_1 and another target keypoint K_3 with respect to preceding K_1 and K_2. We also show examples of masks that ensure causality.

B.2. Causal transformer

Transformer decoder with causal self-attention layers [7, 51] is an alternative architecture that admits autoregressive factorization by construction. We briefly review the causal self-attention layer and analyse its shortcomings in the context of skeleton sequences.

The causal self-attention layer (C-SA) processes each token in a sequence $\mathbf{X}$ by attending only to the preceding tokens. In the case of skeleton keypoint sequences, the corresponding input is an $L \times d$ matrix that collects L d-dimensional tokens. This transformation is parametrized by three learnable projection matrices, $\mathbf{W}_q$, $\mathbf{W}_k$ and $\mathbf{W}_v$ which generate the query, key, and value representations, respectively, as follows:

$$\text{C-SA}(\mathbf{X}) = \left(\sigma(\mathbf{X}\mathbf{W}_q\mathbf{W}_k^\top\mathbf{X}^\top/\sqrt{d}) \odot \mathbf{M}^{\text{SA}}\right)\mathbf{X}\mathbf{W}_v. \tag{22}$$

Here, σ represents the row-wise softmax function, while the mask M ensures causality by not attending the future tokens in the sequence:

$$\mathbf{M}^{\text{SA}}_{ij} = \begin{cases} 1, & \text{if } j \leq i \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$

This design ensures all model parameters are involved in predicting next sequence token.

Our empirical observations indicate that the transformer architecture is a suboptimal design choice for autoregressive density estimation on skeleton sequences.

C. Additional discussion

C.1. On hyperparameter sensitivity.

We tune the SeeKer hyperparameters on the validation subset of UBnormal [3], and apply the same settings to ShanghaiTech and MSAD. We use the early stopping criteria and optimize for maximum validation AUROC. Table 9 reports the performance of SeeKer across key hyperparameters: skeleton sequence length, number of hidden layers, and the input expansion factor that governs the size of the hidden layers. SeeKer consistently achieves competitive performance on the UBnormal validation set across different hyperparameter choices.

Sequence length	AUROC	Nr. hidden layers	AUROC	Expansion factor	AUROC
8	84.9	1	84.8	1	85.4
16	84.1	2	84.8	2	85.4
24	**85.6**	**3**	**85.6**	3	85.6
32	85.5	4	84.6	**4**	**85.6**
48	85.5	5	80.1	5	85.5

Table 9. Validation of the model architecture hyperparameters on the UBnormal validation set shows that minor adjustments in hyperparameter selection have minimal impact on performance.

C.2. On integrating keypoint detection confidence.

Additionally, during training, we train only on confident skeletons. We leverage the per-keypoint detection confidence and define the skeleton confidence as the mean confidence over its keypoints. We filter out skeletons with a detection confidence less than 0.4. This includes occluded skeletons, or skeletons at the border of the frame. Table 10 validates the importance on training only on skeletons with high detection confidence.

Confidence weighting	confidence filtering	UBnormal		ShangahaiTech	
		Full	HR	Full	HR
✗	Eq. (7)	75.5	76.	83.7	84.3
✓	Eq. (8)	**77.9**	**78.9**	**85.5**	**86.9**

Table 10. Training on skeletons with high detection confidence improves the performance in terms of AUROC on UBnormal and ShanghaiTech test.

C.3. On non-human related anomalies.

Figure 9 illustrates SeeKer anomaly scores for an abnormal event from the UBnormal dataset labeled as non-human related (smoke). SeeKer accurately flags the corresponding anomalous frames since people exhibit unusual poses in response to this anomaly. Some works on skeleton-based methods [16, 45] tend to remove non-human related anomalies from the test dataset. However, these test cases highlight the versatility of skeleton-based methods like SeeKer by demonstrating their ability to detect anomalies even when they are only indirectly related to human behavior. Moreover, such testing scenarios regularly appear in relevant benchmarks [3, 34, 62].

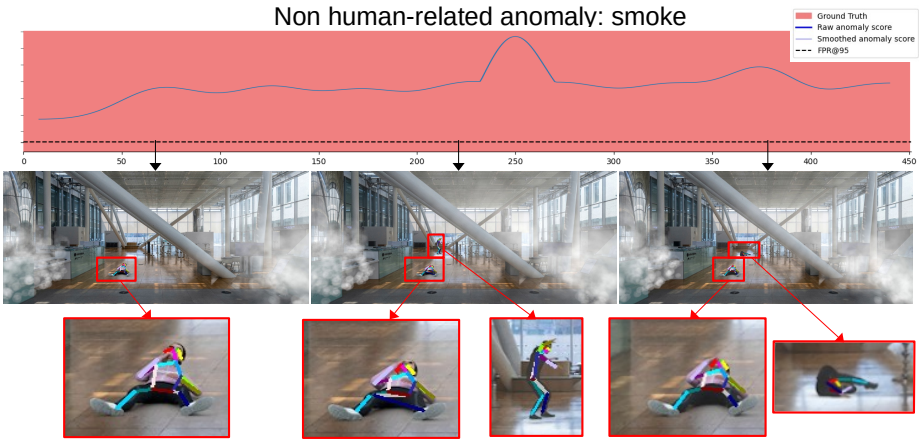

Figure 9. SeeKer can signal non-human related anomalies (e.g. smoke) as long as there are humans in the scene, since humans strike unusual poses in such cases (example from UBnormal).

C.4. On computational requirements.

Our method is efficient in terms of computational requirements. Training requires only 0.76 GB of GPU memory and completes in under 10 minutes on a single NVIDIA GeForce RTX 3090.

D. Limitations

Keypoint extractors. Contemporary skeleton extractors [15] and trackers [54] can struggle in challenging scenarios, such as dense crowds [37] and poor lighting conditions [48]. This may potentially limit the effectiveness of SeeKer in these cases. However, future advances in skeleton extractions can be easily incorporated into our framework and thus enhance SeeKer applicability across diverse conditions.

Potential biases. SeeKer relies on skeleton sequences that are invariant to human-related appearance. Consequently, we cannot introduce any appearance-related biases but may inherit biases of skeleton extractors.