TL;DR: *We propose MoCoDAD that leverages probabilistic diffusion models and conditioning on past motions to accurately detect anomalies by comparing generated motions with expected futures.*

Multimodal Motion Conditioned Diffusion Model for Skeleton-based Video Anomaly Detection

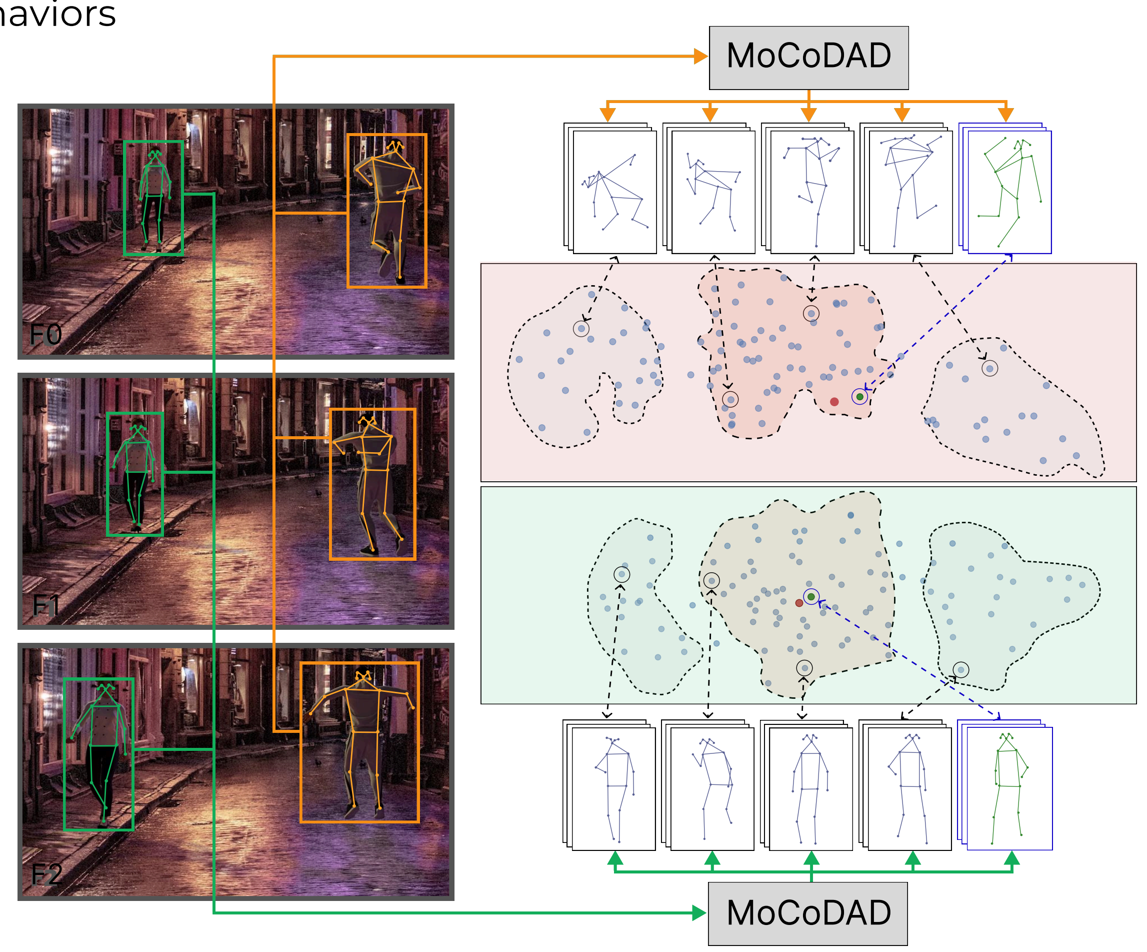
Alessandro Flaborea, Luca Collorone, Guido D'Amely,
Stefano D'Arrigo, Bardh Prenkaj, Fabio Galasso

SAPIENZA UNIVERSITÀ DI ROMA
PINLAB Perception and Intelligence

ICCV23 PARIS

1 Video Anomaly Detection

- Skeleton-based Video Anomaly Detection (VAD)
- Anomalies are rare → learn from regular samples only (**OCC**) or cope with data imbalance
- SoA are constrained to represent a limited latent volume
- Forcing normality into a volume may not work for **diverse-but-still-normal** behaviors

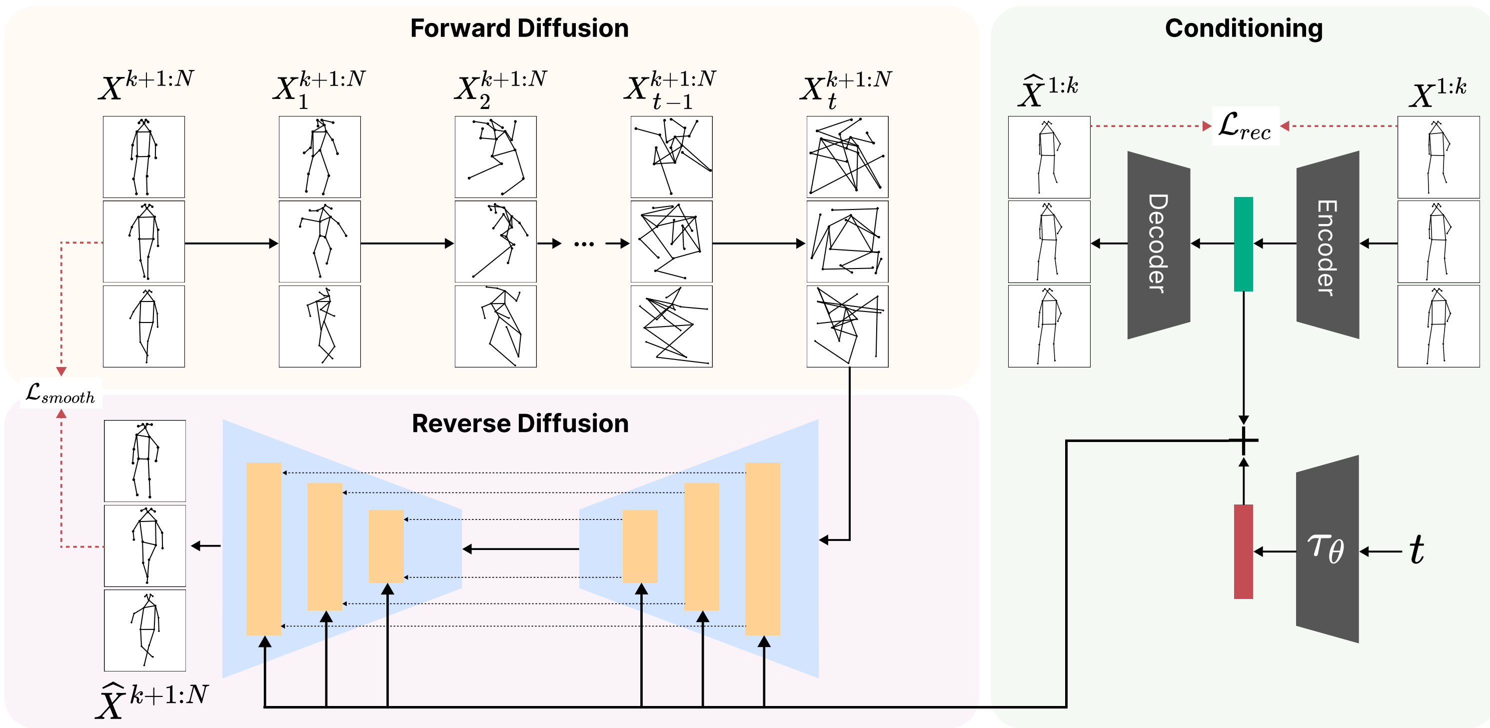


2 How the literature approached VAD

- Latent-based VAD [1,2] scoring data points that fall outside the learned latent space, which represents normality
- Reconstruction-based VAD [3,4] assess how well the model can reconstruct normal input, resulting in higher error rates for anomalies
- Skeleton-based VAD [5-8] methods exploit compact spatio-temporal skeletal representations of human motion instead of raw video frames

[1] M. Sabokrou *et al*. Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. IEEE Transactions on Image Processing, 26(4):1992–2004, 2017
[2] N.T. Nguyen *et al*. Anomaly detection with multiple-hypotheses predictions. In ICML ,pages 4800–4809. PMLR, 2019
[3] W. Liu *et al*. Future frame prediction for anomaly detection–a new baseline. In CVPR, pages 6536–6545, 2018
[4] A. Barbalau *et al*. Ssmtl++: Revisiting self-supervised multi-task learning for video anomaly detection. CVIU, 2023
[5] R. Morais *et al*. Learning regularity in skeleton trajectories for anomaly detection in videos. In CVPR, pages 11996–12004, 2019
[6] W. Luo *et al*. Normal graph:Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection. Neurocomputing, 444:332–337, 2021
[7] A. Markovitz *et al*. Graph embedded pose clustering for anomaly detection. In CVPR, pages 10539–10547, 2020
[8] A. Flaborea *et al*. Contracting Skeletal Kinematic Embeddings for Anomaly Detection. arXiv preprint arXiv:2301.09489, 2023

3 Proposed Approach



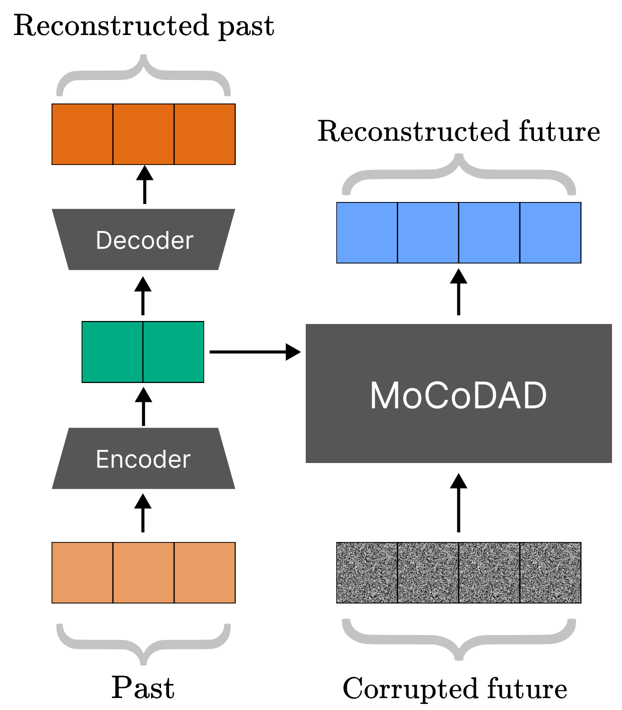
- MoCoDAD learns to reconstruct the **future** corrupted poses by **conditioning on past** poses
- **Training**: *Forward + reverse diffusion process* → the forward process corrupts the coordinates of the joints via a random displacement map
→the reverse process unrolls the corruption via estimating this map

$$\mathcal{L}_{disp} = \mathbb{E}_{t,X,\varepsilon}\left[\left|\left|\varepsilon - \varepsilon_\theta(X_t, t, h)\right|\right|\right] \quad \mathcal{L}_{smooth} = \begin{cases} 0.5 \cdot (\mathcal{L}_{disp})^2 & \text{if } |\mathcal{L}_{disp}| < 1 \\ |\mathcal{L}_{disp}| - 0.5 & \text{otherwise} \end{cases}$$

- **Inference**: Generate multi-modal future sequences of poses from random displacement maps, conditioned on past frames, then aggregates them statistically to detect anomalies
- **Conditioning**: Pass the conditioning of past frames through an encoder, then provide them to all latent layers of the denoising model
- The conditioning embedding adds an auxiliary reconstruction loss

$$\mathcal{L}_{rec} = \left|\left|D(E(X^{1:k})) - X^{1:k}\right|\right|_2^2$$

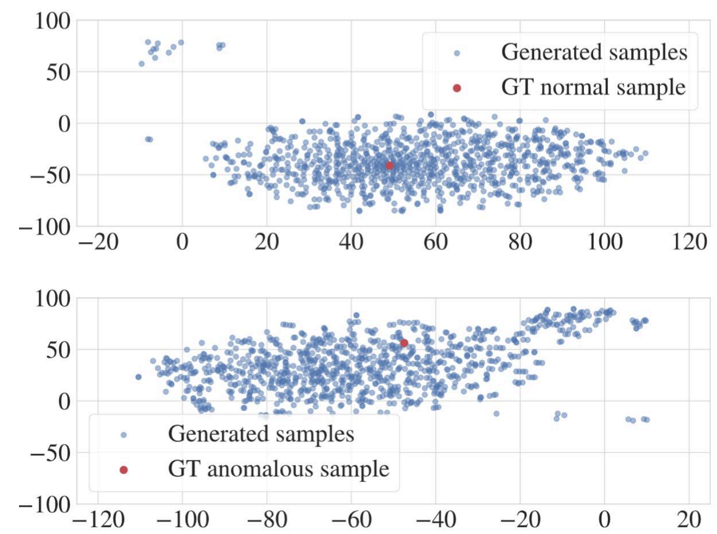
$$\mathcal{L}_{tot} = \lambda_1 \mathcal{L}_{smooth} + \lambda_2 \mathcal{L}_{rec}$$

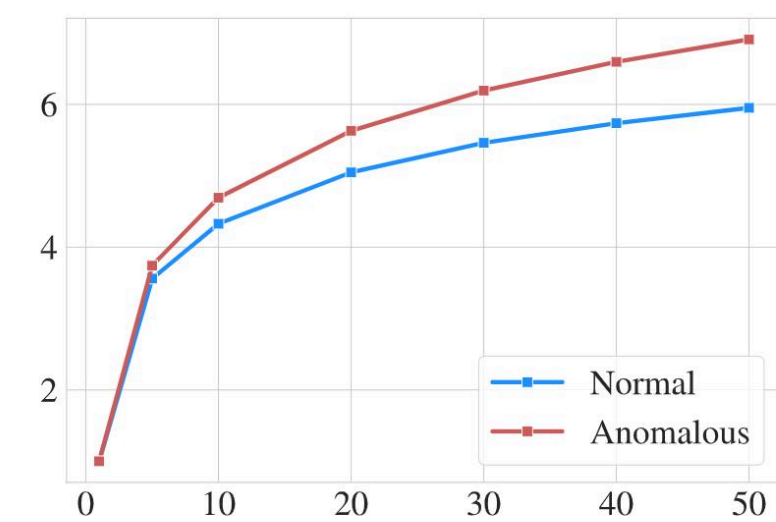


4 Take-away lessons

		HR-STC	HR-Avenue	HR-UBnormal	UBnormal
Conv-AE	*CVPR'16*	69.8	84.8	-	-
Pred	*CVPR'18*	72.7	86.2	-	-
MPED-RNN *	*CVPR'19*	75.4	86.3	61.2	60.6
GEPC *	*CVPR'20*	74.8	58.1	55.2	53.4
Multi-timescale Prediciton *	*WACV'20*	77.0	88.3	-	-
Normal Graph	*Neurocompuring'21*	76.5	87.3	-	-
PoseCVAE *	*ICPR'21*	75.7	87.8	-	-
BiPOCO *	*Arxiv'22*	75.9	87.0	52.3	50.7
STGCAE-LSTM *	*Neurocomputing'22*	77.2	86.3	-	-
SSMTL++	*CVIU'23*	-	-	-	62.1
COSKAD *	*Arxiv'23*	77.1	87.8	65.5	65.0
MoCoDAD *		**77.6**	**89.0**	**68.4**	**68.3**

- Multiple potential futures improve MoCoDAD predictions by reducing penalties on hard-still-normal samples, considered as abnormal by deterministic models
- Normal conditioning motions are centered around the true future; abnormal conditioning makes the ground truth lie on the edge of the predictions' region





- AUC positively correlates with the number of generated future motions for quantiles Q < 0.5, while the correlation is negative for the mean estimate and Q > 0.5