Graph-Jigsaw Conditioned Diffusion Model for Skeleton-based Video Anomaly Detection

Ali Karami[1,2], Thi Kieu Khanh Ho[1,2], Narges Armanfard[1,2]
[1]Department of Electrical and Computer Engineering, McGill University
[2]Mila - Quebec AI Institute, Montreal, QC, Canada
{ali.karami, thi.k.ho, narges.armanfard}@mail.mcgill.ca

This appendix provides supplementary details for the WACV 2025 paper titled *"Graph-Jigsaw Conditioned Diffusion Model for Skeleton-based Video Anomaly Detection"*.

- Sec. 1 provides a comprehensive reviews of related works.
- Sec. 2 presents the pseudocode for our proposed GiCiSAD method.
- Sec. 3 provides the background of diffusion models.
- Sec. 4 presents the results for different statistical aggregations of anomaly scores in the inference phase.
- Sec. 5 provides information of the baselines.
- Sec. 6 provides information of different conditioning strategies.
- Sec. 7 provide a visualization of the *Intra*-community shuffling approach.
- Sec. 8 illustrates an example of the *Inter*-community shuffling approach on the real-world constructed graph.

1. Related Work

1.1. Skeleton-based Video Anomaly Detection

Skeleton-based video anomaly detection (SVAD) has gained significant attention in recent years due to its potential applications in various domains such as video surveillance, healthcare, and human-computer interaction. Many studies have leveraged the powerful representation capabilities of deep learning to automatically learn features from skeleton-based video data, hence, to improve the anomaly detection performance. Existing deep learning studies can be categorized into three main approaches [20, 26]: reconstruction-based, prediction-based and hybrid approaches. In the reconstruction-based approach [7, 22], an autoencoder or its variant model is trained on only normal human activities. During training, the model learns to reconstruct the samples representing normal activities, hence it is expected to yield low reconstruction error for normal data, while achieving high reconstruction error for abnormal data in the test phase. Regarding the prediction-based approach [4, 16, 29], a model is trained to learn the normal human behaviors by predicting the skeletons at the next time steps using information at past time steps. During the test phase, the test samples with high prediction errors are flagged as anomalies. Lastly, the combination of reconstruction-based and prediction-based approaches, which is called as the hybrid approach, has been also widely explored [21, 28, 31]. These methods utilize a multi-objective loss function that consists of reconstruction and prediction errors to learn the characteristics of normal skeletons, aimed at identifying skeletons with large errors as anomalies in the test phase.

However, these three approaches encounter several issues that require more advanced methods to tackle. For example, reconstruction-based methods necessitate the availability of normal data during the training phase, leading to an expectation of higher reconstruction errors for abnormal samples. However, this assumption does not always hold in practice; these methods can also generalize well to anomalies, resulting in false negatives [7]. In prediction-based approaches, determining the optimal prediction horizon for future (or past) events poses a challenge. Moreover, methods relying on future prediction can be sensitive to noise in past data [25]. Even minor alterations in past data can lead to significant variations in predictions, not all of which necessarily indicate anomalies. The combination-based methods include the limitations of the individual learning approaches. It is also challenging to determine the optimum value of combination coefficients (weights) to balance the importance of individual components in a multi-objective loss function. Importantly, in skeleton-based video data,

subtle differences between normal and abnormal actions can oftentimes be localized to specific regions of the body rather than affecting the entire body. However, all existing reconstruction-based, prediction-based and hybrid methods are based on modeling the human body as a whole and ignore the importance of such local variations when detecting anomalies. Note that skeleton-based video data also includes the challenge of inifinite variations of performing normal and abnormal actions. While few studies has addressed this diversity challenge [5], by considering the body as whole, they overlook the fact that abnormalities may be localized to only specific regions of the body, potentially leading to misdetection in cases where anomalies occur in isolated regions while the rest of the body remains normal.

1.2. Self-supervised Learning

Recently, self-supervised learning (SSL) has been widely employed in the context of video anomaly detection [11,12,26]. Essentially, SSL leverages large amounts of unlabeled data to learn meaningful representations without requiring explicit annotations for anomaly detection. Not limited to the predefined reconstruction or prediction tasks, SSL methods define various pretext tasks that can adapt to the specific characteristics and complexities of the data, potentially leading to more robust and discriminative representations. Notable approaches include contrastive learning, which learns to maximize agreement between differently augmented views of the same data, as demonstrated by recent works such as SimCLR [2] and MoCo [8]. Other methods, such as generative adversarial networks (GANs) [1, 12] and autoencoders [13], have also been explored for self-supervised representation learning from videos. While many studies have demonstrated the capability of SSL, they failed to address the challenge of capturing region-specific features in the field of SVAD. Very few works [26] have effectively address this challenge by proposing a challenging pretext task, which encourages the model to focus on region-level features in the image domain. However, it remains unanswered how to adapt this approach to the field of SVAD, particularly considering the presence of skeleton data instead of traditional images in this context. This is due to the fact that unlike images, skeleton data exhibits spatial structure, and the temporal dynamics, which both play a crucial role in defining actions and anomalies. Convolutional layers commonly used in image-based SSL may not directly apply to skeleton data. Instead, architectures based on a combination of recurrent neural networks (RNNs) [18] and graph neural networks (GNNs) [30] can be employed to model the temporal and spatial aspects of skeleton sequences.

1.3. Graph-based Approaches

As denoted in the main paper, skeleton data is inherently a time-series data that exhibits sptatio-temporal dependencies. Hence, it can be naturally represented as graphs [19], where joints correspond to nodes and the connections between joints form edges. Many studies have exploited the potential of graphs for SVAD tasks. For example, [15] introduces a Spatial-temporal Graph Convolutional Autoencoder with Embedded Long Short-Term Memory Network (STGCAE-LSTM) for SVAD. This architecture comprises a single-encoder-dual-decoder setup capable of simultaneously reconstructing the input and predicting future frames. By leveraging graph convolutional operations, the model captures spatial dependencies among joints. However, its fixed adjacency matrix limits its ability to adapt to evolving relationships between joints over time, potentially hindering its performance in capturing dynamic activities. [17] proposes Normal Graph, a spatial-temporal graph convolutional prediction-based network for SVAD. While pioneering in applying graph convolutional networks to SVAD and effectively capturing spatial dependencies, Normal Graph suffers from the same limitation as STGCAE-LSTM in its inability to dynamically learn changing relationships between joints over time, as it fixes the adjacency matrix. Addressing the constraints imposed by fixed adjacency matrices is critical for advancing the state-of-the-art in SVAD. Recent research has explored the capability of dynamically learning graphs overtime in both pure graph and time series domains [3, 10, 27]. In other words, these models dynamically learn the relationships between nodes over time, offering enhanced capabilities in capturing complex spatio-temporal dependencies and detecting anomalies in dynamic graph or time-series domains. However, to date, there remains a scarcity of works capable of effectively capturing the evolving relationships of joints in real-time skeleton-based video streams.

In response to the limitations observed in existing methodologies within the field, we present GiCiSAD, a comprehensive framework that introduces three novel modules to tackle these challenges effectively. The Graph Attention-based Forecasting module leverages a graph learning strategy to effectively capture the spatio-temporal dependencies. To address the issue of region-specific discrepancies, we propose a novel graph-level SSL with a difficult pretext task, called Graph-level Jigsaw Puzzle Maker, which involves various subgraph augmentations applied to the learnable graph, hence providing supervisory signals to help GiCiSAD capture a slight region-level difference between normal and abnormal behaviors. Lastly, to contend with the infinite variations inherent in anomaly detection tasks, GiCiSAD integrates a cutting-edge diffusion-based model named Graph-level Conditional Diffusion Model. Leveraging the learned graph from previous frames as con-

ditional information, this model generates a diverse array of future samples, thereby enhancing the robustness and adaptability of GiCiSAD.

2. GiCiSAD Pseudocode

The overall procedure of the training and inference phases of GiCiSAD is described in Algorithm 1 and Algorithm 2, respectively. Note that during the inference phase, M sets of future frames are generated. Subsequently, these generated frames are compared with the actual ground truth, resulting in M anomaly scores. Finally, these scores are consolidated into a single aggregated value. More detail regarding the aggregation mechanism is presented in Sec. 4. In scenarios with more than one actor in the scene, to summarize the anomaly score of all actors, we follow the methodology outlined in [5]. This approach consolidates the contributions of all actors by considering both the average error across all actors and the span of the error range.

Algorithm 1 GiCiSAD Training

Input: X, diffusion hyperparameters $\{\beta_0, \beta_T, T\}$, δ, λ_1, λ_2, η.
Randomly initialize trainable parameters θ and ψ.
for *not converged* **do**
  $[\mathbf{x}^+, \mathbf{x}^-]$ = Batch(X) ▷ Batching
  Compute $\mathbf{x}^-_{avg}$
  $\mathcal{A}$ = Graph$_\theta(\mathbf{x}^-, \delta)$ ▷ Adjacency Matrix Calculation
  $[\mathcal{A}', p]$ = Puzzle($\mathcal{A}, \eta$) ▷ Puzzle Making
  $[\mathbf{H}, \hat{\mathbf{x}}^-_{avg}]$ = Attention$_\theta(\mathcal{A}')$ ▷ Attention Mechanism
  $\mathcal{H}$ = FC$_\theta(\mathbf{H})$
  $\hat{p}$ = SubgraphHead$_\psi(\mathcal{H})$
  $[t, \mathbf{x}^+_{\text{corrupted}}, \epsilon]$ = Forward($\mathbf{x}^+, T, \beta_0, \beta_T$) ▷ Forward Diffusion
  $\hat{\epsilon}$ = Reverse$_\psi(\mathbf{x}^+_{\text{corrupted}}, \mathcal{H}, f_\psi(t))$ ▷ Reverse Diffusion
  $\mathcal{L} = \lambda_1\left(\mathcal{L}_{\text{graph}}(\hat{\mathbf{x}}^-_{avg}, \mathbf{x}^-_{avg}) + \lambda_2\mathcal{L}_{\text{puzzle}}(\hat{p}, p)\right) + \mathcal{L}_{\text{diffusion}}(\hat{\epsilon}, \epsilon)$
  Backpropage $\mathcal{L}$ to update θ and ψ.
end for

3. Background on Diffusion Models

Diffusion models [9, 23], a class of generative models, define a two-process paradigm that includes: the forward process that slowly adds Gaussian noise to the data and the reverse process that constructs the desired data from the noise. Mathematically, the forward process incrementally adds Gaussian noise to the initial stage, called $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ over T diffusion steps according to a variance scheduler

Algorithm 2 GiCiSAD Inference

Input: $\mathbf{x}^{1:L}$, l, δ, diffusion hyperparameters $\{\beta_0, \beta_T, T\}$, η, M.
Agg $\leftarrow \emptyset$
$\mathbf{x}^-$ = $\mathbf{x}^{1:l}$
$\mathbf{x}^+$ = $\mathbf{x}^{l+1:L}$
Actors = {All Actors participating in $\mathbf{x}^{1:L}$}
for a in Actors **do** ▷ Iteration Over Actors
  Scores $\leftarrow \emptyset$
  for i in range(M) **do** ▷ Generate M Samples
    $\mathbf{u}^+_i = \mathcal{N}(0, \mathbf{I})$
    $\bar{\alpha} = 1$
    for t = T, ..., 1 **do**
      $\mathcal{A}$ = Graph$_\theta(\mathbf{x}^-, \delta)$ ▷ Adjacency Matrix Calculation
      $[\mathcal{A}', p]$ = Puzzle($\mathcal{A}, \eta$) ▷ Puzzle Making
      $\mathbf{H}$ = Attention$_\theta(\mathcal{A}')$ ▷ Attention Mechanism
      $\mathcal{H}$ = FC$_\theta(\mathbf{H})$
      $\hat{\epsilon}$ = Reverse$_\psi(\mathbf{u}^+_i, \mathcal{H}, f_\psi(t))$ ▷ Reverse Diffusion
      $\xi = \mathcal{N}(0, \mathbf{I})$
      $\bar{\alpha} = \bar{\alpha} \times (1 - \beta_t)$
      $\mathbf{u}^+_i = \frac{1}{\sqrt{1-\beta_t}}\left(\mathbf{u}^+_i - \frac{\beta_t}{\sqrt{1-\bar{\alpha}}}\hat{\epsilon}\right) + \xi\sqrt{\beta_t}$ ▷ Recover The Sequence
    end for
    Scores $\leftarrow$ Scores $\cup \{\mathcal{L}_{\text{diffusion}}(\mathbf{x}^+, \mathbf{u}^+_i)\}$ ▷ Save Anomaly Score
  end for
  Agg $\leftarrow$ Agg $\cup$ {AGGREGATE(Scores)} ▷ Aggregate M Anomaly Scores
end for
Anomaly Score: mean(Agg) + log $\frac{1+\max(\text{Agg})}{1+\min(\text{Agg})}$. ▷ Anomaly Score Across All Actors

$\beta_1, \ldots, \beta_T$. The approximate posterior can be represented as:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) := \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \tag{1}$$

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{2}$$

By setting $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$, the forward process allows to immediately transform $\mathbf{x}_0$ to a noisy $\mathbf{x}_t$ according to β_t in a closed form as:

$$q(\mathbf{x}_t|\mathbf{x}_0) := \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I}). \tag{3}$$

The reverse process aims to produce the samples that match the data distribution after a finite number of transition steps. Starting with $p(\mathbf{x}_T) := \mathcal{N}(\mathbf{x}_t; 0, \mathbf{I})$, the joint distribution is then given by:

$$p_\psi(\mathbf{x}_{0:T}) := p(\mathbf{x}_T) \prod_{t=1}^{T} p_\psi(\mathbf{x}_{t-1}|\mathbf{x}_t), \tag{4}$$

$$p_\psi(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\psi(\mathbf{x}_t, t), \sigma_\psi(\mathbf{x}_t, t)). \tag{5}$$

Note that $\mu_\psi(\mathbf{x}_t, t)$ and $\sigma_\psi(\mathbf{x}_t, t)$ are parameterized as:

$$\mu_\psi(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \Big(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\psi(\mathbf{x}_t, t) \Big), \tag{6}$$

$$\sigma_\psi(\mathbf{x}_t, t) = \sqrt{\bar{\beta}_t}, \tag{7}$$

where $\bar{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$, and $\bar{\beta}_1 = \beta_1$. ϵ_ψ is a network approximator (the U-Net-based architecture in our case), which take $\mathbf{x}_t$ and the diffusion step t as the inputs, and aims to predict the noise from $\mathbf{x}_t$.

4. Different Strategies for Statistical Aggregations

In this analysis, we assess the performance of anomaly detection by altering the method of aggregation. Given the infinite variations in executing both normal and abnormal actions, we generate M sets of future frames. For each set, we calculate an anomaly score. As discussed in the main paper, for the purpose of statistically aggregating these scores, we explore four strategies: taking the mean, the median, the maximum distance, and the minimum distance. In the mean and median approaches, we derive either the mean or the median of all M scores and allocate this value to the respective frame to evaluate its anomaly level. Regarding the maximum and minimum distance selector approach, the highest and lowest anomaly score among all scores is assigned to the frame respectively. Comparison between these four methods is shown in Tab. 1, with the minimum distance approach demonstrating superior performance across the board. The suboptiomal performance of the maximum distance strategy further supports the idea that generated future samples that are conditioned on normal motions are as diverse as those that are conditioned on anomalous motions. This is due to the fact that if normal conditioned future sampled were not diverse, both the maximum and minimum distance strategies would have resulted in identical outcomes. Fig. 1 further demonstrates the effectiveness of our proposed GiCiSAD method in generating a diverse range of samples conditioned on both normal and abnormal frames. As can be seen, when the model is conditioned on normal past frames, the generated future frames are diverse yet close to the ground truth, with low anomaly scores.

Aggregation Strategy	**HR-Avenue**	**HR-STC**
Mean	89.5	77.8
Median	89.5	77.9
Maximum Distance	88.2	77.3
Minimum Distance	**89.6**	**78**

Table 1. Comparison between different aggregation strategies for 50 generation of future frames, assessed through the AUROC metric on the HR-Avenue and HR-STC datasets.

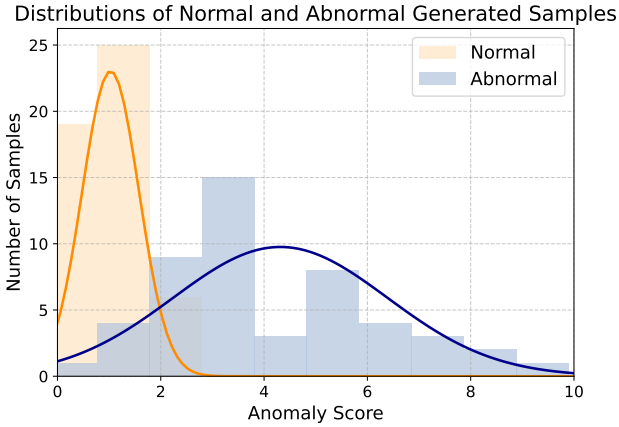

Figure 1. Histograms of the anomaly scores for 50 future frames generated by `Diffusion` on the HR-STC dataset, for both cases of conditioning on normal and abnormal past frames.

Conversely, when conditioned on abnormal past frames, the generated frames remain diverse but deviate significantly from the ground truth.

5. Baselines

As mentioned in the main paper, we compare GiCiSAD against SOTA methods. Details of each method are described below.

1. GEPC [19] analyzes human poses through graphs. By mapping these graphs into a latent space and clustering them, they represent each action based on its soft-assignments to these clusters, akin to a "bag of words" model where actions are defined by their resemblance to foundational action-words. They then employ a Dirichlet process-based mixture model to classify actions as normal or anomalous.

2. PoseCVAE [14] predicts future pose trajectories based on a sequence of past normal poses, aiming to learn a conditional posterior distribution that characterizes the normal data, using a conditional variational autoencoder. They also propose a self-supervised component

to enhance the encoder and decoder's ability to capture the latent space representations of human pose trajectories effectively. They imitate abnormal poses in the embeded space and use a binary cross-entropy loss along with the standard conditional variational autoencoder loss function.

3. STGCN-LSTM [15] merges spatial-temporal graph convolutional autoencoder and Long Short-term Memory networks. They use reconstruction and future prediction errors for detecting anomalies.

4. COSKAD [6] utilizes a graph convolutional network to encode skeletal human motions, and learns to project skeletal kinematic embeddings onto a latent hypersphere of minimal volume for video anomaly detection. COSKAD innovates by proposing three types of latent spaces: the traditional Euclidean, their proposed spherical and hyperbolic spaces.

5. MoCoDAD [5] utilizes autoencoder conditioned diffusion probabilistic models to generate a variety of future human poses. Their autoencoder-based approach conditions on individuals' past movements and leverages the enhanced mode coverage of diffusion processes to produce diverse yet plausible future motions. By statistically aggregating these potential futures, the model identifies anomalies when the forecasted set of motions diverges significantly from the observed future.

6. TrajREC [24] leverages multitask learning to encode temporally occluded trajectories, jointly learn latent representations of the occluded segments, and reconstruct trajectories based on expected motions across different temporal segments.

6. Weaker Forms of Conditioning Mechanism

This section elaborates on the *Encoder*-based and *AutoEncoder*-based conditioning mechanisms [5] that are used for comparison with our proposed *Graph*-based approach, mentioned in the ablation study of the main paper. The objective of conditioning mechanism is to generate an efficient latent representation of past frames, $\mathcal{H}$, to effectively guide the `Diffusion` process. The architecture of these two conditioning mechanisms is illustrated in Fig. 2. $\mathcal{H}$ will be used as the conditioning signal to guide the `Diffusion`, where the architecture of `Diffusion` remains unchanged. The *Encoder*-based method introduces no additional loss to the model. Conversely, the *AutoEncoder*-based approach incorporates the reconstruction loss of the past frames into the `Diffusion` loss, thereby modifying the overall loss calculation as follows.

$$\mathcal{L} = \lambda \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{diffusion}}, \tag{8}$$

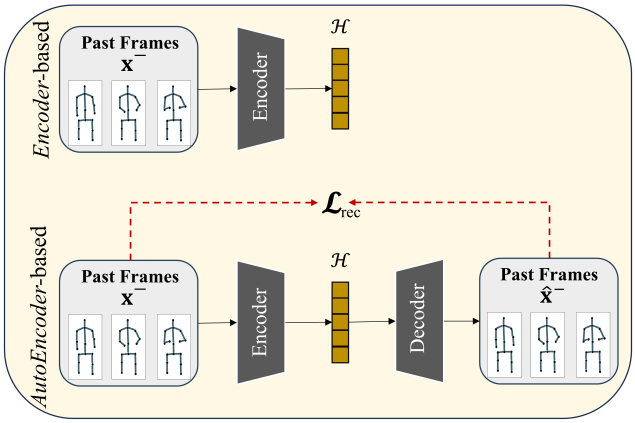

Figure 2. Comparison of *Encoder*-based and *Autoencoder*-based conditioning mechanisms.

where λ is 0.1. From an architectural perspective, the encoder features a channel sequence of $(32, 16, 32)$, incorporating a bottleneck dimension of 32 and a latent projector with a dimensionality of 16.

7. Visualization of *Intra*-Community Shuffling Approach

The visualization of the *Intra*-Community shuffling approach is shown in Fig. 3. A detailed description of this approach has previously been provided in the ablation study section, "Types of Graph-based Jigsaw Puzzles," of the main paper.

8. Visualization of *Inter*-Community Shuffling Approach on Real Constructed Graphs

While we have presented a simple and easy-to-understand visualization of our *Inter*-community shuffling approach in Fig. 2 of the main paper, we provide Fig. 4 for a more detailed view of the *Inter*-community shuffling process on the real constructed graphs. Graphs include 34 nodes (joints), δ and η are set to 4. Note that the graphs (before and after shuffling) are directed, indicating that connections are not inherently symmetric. In this figure, Subgraph 2 (depicted in green) is shuffled with Subgraph 1 (depicted in orange). Specifically, the densest nodes of Subgraph 2, namely, $\{32, 3, 31, 30, 7, 13\}$, are shuffled with nodes $\{29, 4, 6, 11, 26, 2\}$ from Subgraph 1, respectively. After the shuffling process, while nodes of the smaller subgraph, i.e., Subgraph 1, stay connected, the intra-connections of the larger subgraph, i.e., Subgraph 2, undergoes significant changes, nearly dividing it into two distinct parts. It should be noted that the other two subgraphs, i.e., Subgraphs 0 and 3, retain their connections, while only their spatial positioning is changed for better visualization in the figure.

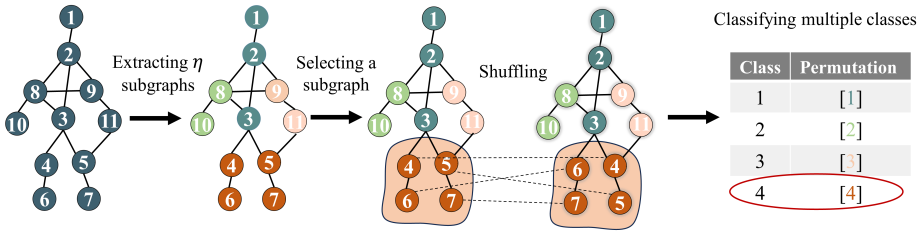

Figure 3. Visualization of the *Intra*-Community shuffling approach. Nodes with the same color formulate a subgraph. Note that although each node is required to have δ connections, for improved visualization, this property is not strictly maintained in the figure.

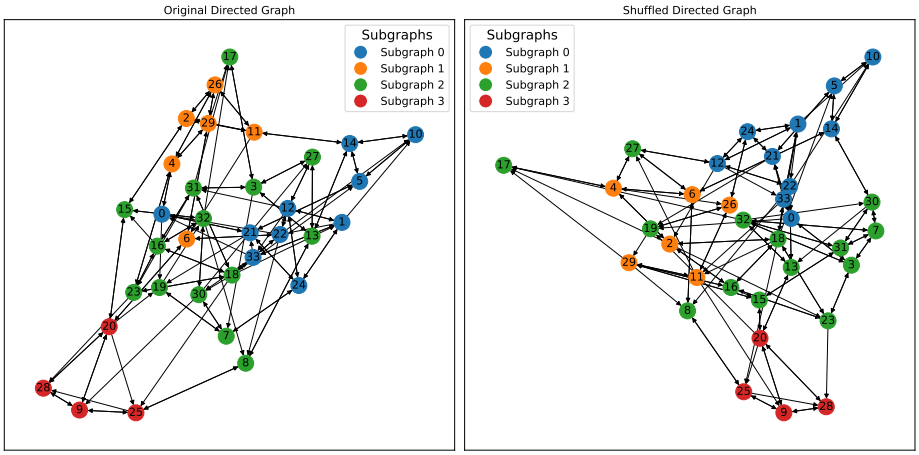

Figure 4. *Inter*-Community shuffling process between Subgraph 2 and Subgraph 1.

References

[1] Dongyue Chen, Lingyi Yue, Xingya Chang, Ming Xu, and Tong Jia. Nm-gan: Noise-modulated generative adversarial network for video anomaly detection. *Pattern Recognition*, 116:107969, 2021. 2

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[3] Ailin Deng and Bryan Hooi. Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4027–4035, 2021. 2

[4] Xinyang Feng, Dongjin Song, Yuncong Chen, Zhengzhang Chen, Jingchao Ni, and Haifeng Chen. Convolutional transformer based dual discriminator generative adversarial networks for video anomaly detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5546–5554, 2021. 1

[5] Alessandro Flaborea, Luca Collorone, Guido Maria D'Amely Di Melendugno, Stefano D'Arrigo, Bardh Prenkaj, and Fabio Galasso. Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10318–10329, 2023. 2, 3, 5

[6] Alessandro Flaborea, Guido Maria D'Amely di Melendugno, Stefano D'arrigo, Marco Aurelio Sterpa, Alessio Sampieri, and Fabio Galasso. Contracting skeletal kinematic embeddings for anomaly detection. *arXiv preprint arXiv:2301.09489*, 2023. 5

[7] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1705–1714, 2019. 1

[8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[10] Thi Kieu Khanh Ho, Ali Karami, and Narges Armanfard. Graph-based time-series anomaly detection: A survey and outlook. *arXiv preprint arXiv:2302.00058*, 2024. 2

[11] Liang Hu, Dora D Liu, Qi Zhang, Usman Naseem, and Zhong Yuan Lai. Self-supervised learning for multilevel skeleton-based forgery detection via temporal-causal consis-

tency of actions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 844–853, 2023. 2

[12] Chao Huang, Jie Wen, Yong Xu, Qiuping Jiang, Jian Yang, Yaowei Wang, and David Zhang. Self-supervised attentive generative adversarial networks for video anomaly detection. *IEEE transactions on neural networks and learning systems*, 2022. 2

[13] Chao Huang, Zehua Yang, Jie Wen, Yong Xu, Qiuping Jiang, Jian Yang, and Yaowei Wang. Self-supervision-augmented deep autoencoder for unsupervised visual anomaly detection. *IEEE Transactions on Cybernetics*, 52(12):13834–13847, 2021. 2

[14] Yashswi Jain, Ashvini Kumar Sharma, Rajbabu Velmurugan, and Biplab Banerjee. Posecvae: Anomalous human activity detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2927–2934. IEEE, 2021. 4

[15] Nanjun Li, Faliang Chang, and Chunsheng Liu. Human-related anomalous event detection via spatial-temporal graph convolutional autoencoder with embedded long short-term memory network. *Neurocomputing*, 490:482–494, 2022. 2, 5

[16] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6536–6545, 2018. 1

[17] Weixin Luo, Wen Liu, and Shenghua Gao. Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection. *Neurocomputing*, 444:332–337, 2021. 2

[18] Weixin Luo, Wen Liu, Dongze Lian, Jinhui Tang, Lixin Duan, Xi Peng, and Shenghua Gao. Video anomaly detection with sparse coding inspired deep neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):1070–1084, 2019. 2

[19] Amir Markovitz, Gilad Sharir, Itamar Friedman, Lihi Zelnik-Manor, and Shai Avidan. Graph embedded pose clustering for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10539–10547, 2020. 2, 4

[20] Pratik K Mishra, Alex Mihailidis, and Shehroz S Khan. Skeletal video anomaly detection using deep learning: Survey, challenges, and future directions. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024. 1

[21] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11996–12004, 2019. 1

[22] Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1273–1283, 2019. 1

[23] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[24] Alexandros Stergiou, Brent De Weerdt, and Nikos Deligiannis. Holistic representation learning for multitask trajectory anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6729–6739, 2024. 5

[25] Yao Tang, Lin Zhao, Shanshan Zhang, Chen Gong, Guangyu Li, and Jian Yang. Integrating prediction and reconstruction for anomaly detection. *Pattern Recognition Letters*, 129:123–130, 2020. 1

[26] Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *European Conference on Computer Vision*, pages 494–511. Springer, 2022. 1, 2

[27] Feng Xia, Ke Sun, Shuo Yu, Abdul Aziz, Liangtian Wan, Shirui Pan, and Huan Liu. Graph learning: A survey. *IEEE Transactions on Artificial Intelligence*, 2(2):109–127, 2021. 2

[28] Muchao Ye, Xiaojiang Peng, Weihao Gan, Wei Wu, and Yu Qiao. Anopcn: Video anomaly detection via deep predictive coding network. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1805–1813, 2019. 1

[29] Guang Yu, Siqi Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin, and Marius Kloft. Cloze test helps: Effective video anomaly detection via learning to complete video events. In *Proceedings of the 28th ACM international conference on multimedia*, pages 583–591, 2020. 1

[30] Xianlin Zeng, Yalong Jiang, Wenrui Ding, Hongguang Li, Yafeng Hao, and Zifeng Qiu. A hierarchical spatio-temporal graph convolutional neural network for anomaly detection in videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(1):200–212, 2021. 2

[31] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1933–1941, 2017. 1