SCD-Net: Spatiotemporal Clues Disentanglement Network for Self-Supervised Skeleton-Based Action Recognition

Cong Wu\textsuperscript{1,2}, Xiao-Jun Wu\textsuperscript{1*}, Josef Kittler\textsuperscript{2}, Tianyang Xu\textsuperscript{1}, Sara Ahmed\textsuperscript{2}, Muhammad Awais\textsuperscript{2}, Zhenhua Feng\textsuperscript{2}

\textsuperscript{1}School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China
\textsuperscript{2}School of Computer Science and Electronic Engineering, and CVSSP, University of Surrey, Guildford GU2 7XH, UK
congwu@stu.jiangnan.edu.cn, wu_xiaojun@jiangnan.edu.cn; tianyang_xu@163.com; \{sara.atito, muhammad.awais, j.kittler, z.feng\}@surrey.ac.uk

Abstract

Contrastive learning has achieved great success in skeleton-based action recognition. However, most existing approaches encode the skeleton sequences as entangled spatiotemporal representations and confuse the contrasts to the same level of representation. Instead, this paper introduces a novel contrastive learning framework, namely Spatiotemporal Clues Disentanglement Network (SCD-Net). Specifically, we integrate the decoupling module with a feature extractor to derive explicit clues from spatial and temporal domains respectively. As for the training of SCD-Net, with a constructed global anchor, we encourage the interaction between the anchor and extracted clues. Further, we propose a new masking strategy with structural constraints to strengthen the contextual associations, leveraging the latest development from masked image modelling into the proposed SCD-Net. We conduct extensive evaluations on the NTU-RGB+D (60&120) and PKU-MMD (I&II) datasets, covering various downstream tasks such as action recognition, action retrieval, transfer learning, and semi-supervised learning. The experimental results demonstrate the effectiveness of our method, which outperforms the existing state-of-the-art (SOTA) approaches significantly. Our code and supplementary material can be found at https://github.com/cong-wu/SCD-Net.

1 Introduction

Skeleton-based action recognition focuses on identifying human actions via skeleton sequences, which has witnessed significant advancements in recent years. On one hand, deep networks, such as Graph Convolutional Network (GCN) (Yan, Xiong, and Lin 2018), have been investigated and successfully applied for the task at hand. On the other hand, several large-scale datasets, \textit{e.g.}, NTU-\textit{RGB+D} (Shahroudy et al. 2016), have been proposed, providing an experimental foundation for further development of the area. However, like most visual tasks, the training of a high-performance model typically requires a massive amount of high-quality labelled data. This requirement poses a significant challenge in data collection and annotation. Fortunately, self-supervised learning has emerged as a solution to address this challenge by leveraging inherent associations instead of relying on annotations. In particular, recent investigations (Dong et al. 2023) have demonstrated that contrastive learning, owing to its interpretability and transferability, has emerged as a front-runner in self-supervised skeleton-based action recognition.

However, several crucial aspects are disregarded by existing approaches. First, the encoder is responsible for mapping the input into a latent space where the contrast can be conducted. While most previous methods (Zhang et al. 2022a; Franco et al. 2023) concentrate on obtaining unified information through commonly used spatiotemporal modelling networks. Their designs result in the complete entanglement of information, failing to provide clear indications for subsequent contrastive measures. There have been sporadic attempts (Dong et al. 2023) aiming to extract absolutely isolated spatial or temporal information. But repeated evidence has shown that complete isolation of spatiotemporal information is suboptimal for action recognition (Kay et al. 2017; Lin, Gan, and Han 2019). More importantly, most approaches focus on constructing contrast pairs at same level of representation (Guo et al. 2022) during optimisation; Or attempt to force the interaction between information flows, overlooking the gap between domains (Dong et al. 2023). In addition, existing techniques (Thoker et al. 2021) often limit themselves to scale transformation, which results in not fully capitalising on the potential of data augmentation. Here we introduce a novel contrastive learning framework that focuses on disentangling spatiotemporal clues, and exploits masking in data augmentation to provide more discriminative inputs, thereby prompting the model to learn more robust interactions.

To leverage the intricate features present in skeleton sequences, we propose a dual-path decoupling encoder to generate explicit representations from spatial and temporal domains. Our encoder comprises two main subsystems: a feature extractor and a decoupling module. The role of the feature extractor is to extract fundamental spatiotemporal features from skeleton sequences as the intermediate representations. Since lacking an overall grasp of the skeleton sequence, it is difficult to obtain a picture of the features simply by modelling from a certain perspective. Next, we generate token embeddings by projection and refine the sequence features with a transformer-based module. The decoupling modules are instrumental to deriving disentangled
techniques were employed to classify skeletons via traditional feature extraction methods. Recently, GCN-based approaches (Yan, Xiong, and Lin 2018; Li et al. 2019; Liu et al. 2020) have gained prominence in the field. The general paradigm initially models the skeleton sequence as a spatiotemporal graph and subsequently employs information aggregation and updating techniques. Inspired by the notable achievements of transformer (Dosovitskiy et al. 2020; Liu et al. 2022), some recent methods (Zhang et al. 2021, 2022b) have explored its powerful sequence modelling capability for skeleton-based tasks.

2.2 Contrastive Learning

Contrastive learning is a typical solution for self-supervised learning. Unlike generative learning (Zhu et al. 2020; Huang et al. 2022), contrastive learning does not involve explicit generation or reconstruction of the input. Instead, it focuses on learning discriminative representations through a contrastive loss. Most contrastive learning methods (Chen et al. 2020; Grill et al. 2020) operate on the principle of pulling positive pairs closer to each other, while simultaneously pushing dissimilar pairs farther apart within a projection space. By exploring the internal properties within the data, contrastive learning enables learning more generalised and robust representations, resulting in remarkable performance on downstream tasks (Wang and Liu 2022).

2.3 Contrastive Learning for Skeleton-based Action Recognition

Contrastive learning has also been successfully employed in skeleton-based action recognition. Thoker et al. (Thoker et al. 2021) proposed the intra-skeleton and inter-skeleton contrastive loss, achieving promising results in several downstream tasks. Dong et al. (Dong et al. 2023) utilised down-sampling operations at different stages of the encoder to obtain multi-scale features for constructing a hierarchical contrastive learning framework. Franco et al. (Franco et al. 2023) proposed a novel approach that involves projecting the encoded features into a hyperbolic space, which is a non-Euclidean space that allows more efficient modelling of complex association. Despite these advances, most existing studies overlook the crucial step of extracting and disentangling spatial and temporal clues from skeleton sequences, not to mention the failure of considering the interactions among representations of different domains.

For contrastive learning, data augmentation processes the training sample to obtain positive input pairs with certain differences. Thoker et al. (Thoker et al. 2021) used various spatiotemporal augmentation techniques, including pose augmentation, joint jittering, and temporal crop-resize, to generate different inputs for the query and key encoders. While most methods follow similar scale transformation paradigms, Zhou et al. (Zhou et al. 2023) proposed a strategy of masking selected nodes and frames, which greatly extends the augmentation to "destroy" the data structure. However, unlike image data, skeleton sequences have strong physical associations, meaning that even if a certain node or frame is corrupted, it can easily be corrected using the information from adjacent areas (Cheng et al. 2020). Incorporating the structural constraints, we expand the point-based

![Figure 1: A comparison of the proposed method with HiCo-Transformer (Dong et al. 2023), using multiple evaluation metrics. (Better view in colour.)](image-url)
masking approach to area-based masking. This extension aims to prevent potential data leakage and enhance the learning capabilities of SCD-Net.

### 3 The Proposed SCD-Net

In this section, we will initially present the overall framework of SCD-Net, followed by a detailed introduction to each of its components in the subsequent sections.

#### 3.1 The Overall Framework

The overall pipeline of the proposed method consists of two branches, as shown in Figure 2 (b). Each branch has the same components, including data augmentation and encoder. For any input data, we link the outputs obtained by the encoder and momentum encoder to form contrast pairs.

To elaborate further, the input of the network is defined as a sequence of human body key points, denoted as $\mathcal{X} \in \mathbb{R}^{C \times T \times V}$, where $T$ is the length of the sequence, $C$ is the physical coordinate defined in a 2D/3D space, $V$ is the number of key points. In SCD-Net, we first apply data augmentation to generate the augmented views for the encoders. Second, for each encoder, we deploy feature extraction and (spatial/temporal) decoupling operations to generate spatial feature $z_s \in \mathbb{R}^{C_2}$ and temporal feature $z_t \in \mathbb{R}^{C_2}$ from the entangled information. Third, we project these clues into the same semantic space to obtain the final representations.

The loss function, $L_{\theta, \xi}$, is defined as a measure of interactions of these representations. The parameters $\theta$ and $\xi$ specify the architecture corresponding to the encoder and the momentum encoder. During the optimisation, the loss is backpropagated through the model, while the parameters of the momentum encoder are updated using a momentum update strategy. So the final optimiser is:

$$ \theta \leftarrow \text{optimizer}(\theta, \nabla_\theta L_{\theta, \xi}, \tau), \quad \xi = \xi * m + \theta * (1 - m), \quad (1) $$

where $r$ and $m$ are the learning rate and decay rate.

#### 3.2 The Dual-path Decoupling Encoder

In general, the features extracted from a skeleton sequence are characterised as complex spatiotemporal associations describing an action. However, we argue that this paradigm is not suitable for contrastive learning. As the information is greatly entangled, it is difficult to provide clear guidance for the subsequent comparison. In SCD-Net, we advocate a dual-path decoupling encoder to extract clear and multiple discriminative cues from the complex sequence information. Such clues provide clear instructions for a subsequent contrast quantification. More importantly, a reliable assessment of the contrast between different domains is likely to provide stronger discrimination.

For brevity, we generally denote the augmented input for the encoder as $\mathcal{X}$. As demonstrated by the existing studies, completely isolating the information flow is suboptimal (Lin, Gan, and Han 2019; Wang et al. 2021). Given that, we apply a spatiotemporal modelling network to extract the intermediate features. Inspired by the excellent performance in modelling skeleton sequences (Yan, Xiong, and Lin 2018), we use a $l_2$-layer GCN, consisting of spatial-GCN (S-GCN) and temporal GCN (T-GCN), to obtain unified representations $\mathcal{Y} \in \mathbb{R}^{C_1 \times T \times V}$. This can be expressed as a process of aggregation and updating of adjacent features. Specifically, for any $\mathcal{X}_{ti} \in \mathbb{R}^{C_1}$, where $t$ and $i$ are the frame and joint index, the newly generated features $\mathcal{Y}_{ti} \in \mathbb{R}^{C_1}$, can be expressed as:

$$ \mathcal{Y}_{ti} = \sum_{\mathcal{X}_{uj} \in B(\mathcal{X}_{ti})} \frac{1}{Z_{ti}(\mathcal{X}_{uj})} \cdot \mathcal{X}_{uj} \cdot w(1_{ti}(\mathcal{X}_{uj})), \quad (2) $$

where $B(\mathcal{X}_{ti})$ denotes the kernel of the graph convolution operation on $\mathcal{X}_{ti}$, $Z(\cdot)$ represents normalisation, $w(\cdot)$ is the weight function, and $1(\cdot)$ maps adjacent nodes to the corresponding subset index.

Given the intermediate spatiotemporal representation $\mathcal{Y}$, the following step is decoupling operation, which involves projection and refinement, as shown in Figure 3. Specifically, we perform a dimension transformation on $\mathcal{Y}$ to derive $\mathcal{Y}_{rt} \in \mathbb{R}^{C_1 \times C_1 T}$ and $\mathcal{Y}_{rt} \in \mathbb{R}^{C_1 \times C_1 V}$. These transformed representations are then projected to higher semantic space to obtain the corresponding spatial and temporal

Figure 2: Our model benefits from three innovations: a dual-path encoder for distinct spatiotemporal information decoupling; a bespoke cross-domains contrastive loss promoting the information interaction; a structurally-constrained masking strategy for efficient data augmentation.

Figure 3: The dual-path decoupling module that provides clean spatial and temporal representations of a skeleton sequence.
The spatial embedding operation is defined as:
\[
Y_s = W_{s2} \ast \text{ReLU}(W_{s1} \ast Y_{rs} + B_{s1}) + B_{s2},
\]
where \(W\) and \(B\) are the trainable weights and bias, \(Y_s \in \mathbb{R}^{V \times C_2}\). However, the current embedding is still a rough representation as the current features lack explicit interactions within points or frames. While the feature extraction operation incorporates significant spatiotemporal interactions, these interactions often become intertwined. Hence, it remains crucial to address the interaction of individual spatial and temporal representations. Here we use a \(l_t\)-layer self-attention network to construct the self-correlation information extraction process that refines the spatial and temporal representations, as shown in Figure 3. The transformer architecture used in this method has two main components: self-attention and feed-forward modules. For instance, we obtain \(z_s \in \mathbb{R}^{C_2}\) as follows:
\[
\hat{Z}_{si} = \text{SoftMax}\left(\frac{F_{qi}(Y_s) \cdot (F_{ki}(Y_s))^T}{\sqrt{d_i}}\right) \cdot (F_{vi}(Y_s)),
\]
\[
\hat{Z}_s = \text{LN}(\text{Concat}([\hat{Z}_{s1}, ..., \hat{Z}_{si}, ..., \hat{Z}_{sh}]) + Y_s),
\]
where \(F\) represents feature projection, implemented by fully connected layers, with Concat denoting the concatenation operation, \(\text{LN}\) and \(\text{FFN}\) means layer normalisation and Feed-Forward Networks, \(h\) signifies the number of heads. \(z_t \in \mathbb{R}^{C_2}\) can also be obtained by similar operations.

### 3.3 Cross-domain Contrastive Loss

With the decoupled spatial and temporal representations, as shown in Figure 2, we first obtain the final representations by:
\[
q_s = F_s(z_s), \quad q_t = F_t(z_t),
\]
where \(F_s, F_t\) are the corresponding projection functions, which can be defined by two fully connected layers, similar to Eq. (3). As we discussed earlier, there is an obvious gap between spatial and temporal domains, for which we introduce a global perspective \(q_g\) compatible with both as a intermediary for contrasts.
\[
q_g = F_g[z_t, z_s],
\]
where \(F_g\) are the corresponding projection function. The outputs \((k_s, k_t, k_g)\) of the corresponding key encoder, can also be obtained by a similar process.

Based on these candidate features, we define a new cross-domain loss. The core of our design lies in anchoring the global representation and building its association with other representations obtained by another encoder. The loss function is defined as:
\[
\mathcal{L}_{g,\xi} = \lambda_1 \cdot \mathcal{L}(q_s, k_s) + \lambda_2 \cdot \mathcal{L}(q_t, k_t) + \lambda_3 \cdot \mathcal{L}(q_g, k_g) + \lambda_4 \cdot \mathcal{L}(q_t, k_g),
\]
where \(\lambda\) is the mixing weight of the sum operation. Specifically, for any given contrast pair \(u\) and \(v\), \(\mathcal{L}(u, v)\) evaluates the correlation between \(u\) and \(v\). The objective is to minimise the distance between positive pairs from the query and key encoders, while maximising the distance from the other features.

To achieve this, we employ the contrastive loss based on InfoNCE (Oord, Li, and Vinyals 2018) as follows:
\[
\mathcal{L}(u, v) = -\log \frac{h(u, v)}{h(u, v) + \sum_{m \in M} h(u, m)},
\]
where \(h(u, v) = \exp(u \cdot v / r)\) is the exponential similarity measurement. We denote the first-in-first-out queue of the previously extracted features, containing \(l_m\) negative samples, by \(M\).

### 3.4 Data Augmentation

By imposing structural constraints, our approach applies the masking operation within a local region around the current randomly selected joints or frames instead of relying on isolated points or frames. In this way, we substantially eliminate explicit local contextual associations, and force the encoders to model robust contextual relationships through interactive contrastive learning.

#### Structurally Guided Spatial Masking

Considering the physical structure of skeleton, when a certain joint is selected for masking, we simultaneously mask the points in its adjacent area. Let us represent the adjacency relationship using the matrix \(P\). \(P_{ij} = 1\), if joints \(i\) and \(j\) are connected, otherwise \(P_{ij} = 0\). We denote \(D = \mathbb{P}^n\), where \(n\) is the exponent. The element \(D_{ij}\) in \(D\) represents the number of paths that can be taken to reach node \(j\) from node \(i\) by walking \(n\) steps. Note that reversal and looping are allowed. To impose a structural constraint, when node \(i\) is selected, we perform the same augmentation operation on all nodes \(j\) for which \(D_{ij} \neq 0\). The only undesirable artefact of this operation is that it may give rise to a variable number of candidate joints. To avoid this, for several randomly selected nodes, the actual augmentation is applied only to a fixed number \((k)\) of points exhibiting the highest overall response on \(D\).

#### Cube-based Temporal Masking

The sequence follows a linear relationship in time. To avoid information leakage between adjacent frames (Tong et al. 2022), we construct a cube, defined by a selected segment and its adjacent frames. Specifically, we start by dividing the input sequence into cubes as shown in Figure 4. We randomly select \(r\) cubes of equal length. Next, we randomly select \(s\) cubes as candidates for masking.

#### 4 Experiments

#### 4.1 Experimental Settings

**Datasets** We evaluate the proposed method on four benchmarking datasets, NTU-RGB+D (60&120) (Shahiroudy et al. 2016) and PKU-MMD (I&II) (Liu et al. 2017).
### Implementation Details

For the input data, 64 frames are randomly selected for training and evaluation. We perform data augmentation operations, including rotate, flip and shear, as well as the proposed structural spatial masking and temporal masking, on the selected sequence. Each operation has a 50% chance of being executed. For masking, we set $n = 2$, $k = 8$, $s = 16$, $r = 6$. For the encoder, we refer to MoCo (He et al. 2020) and build a query encoder and the corresponding key encoder. The two encoders have exactly the same structure as shown in Figure 3. For feature extractor, we borrow the structure from CTR-GCN (Chen et al. 2021) as the basic operation. For network optimisation, we set the queue length of $M$ to 8192 (except 2048 for PKU-MMD I), moco momentum to 0.999, softmax temperature to 0.2, and $\lambda$ to 1.

More details are presented in supplementary material.

### 4.2 Comparison with the SOTA Methods

We compare SCD-Net with several SOTA methods, including: (1) Encoder-decoder based methods: LongT GAN (Zheng et al. 2018), EnGAN-PoseRNN (Kundu et al. 2019), H-Transformer (Cheng et al. 2021), SeBiReNet (Nie, Liu, and Liu 2020), Colorization (Yang et al. 2021), GL-Transformer (Kim et al. 2022); (2) Hybrid learning based methods: ASSL (Si et al. 2020), MS2L (Lin et al. 2020), PCRP (Xu et al. 2021), HiTRS (Chen et al. 2022); (3) Contrastive-learning based methods: CrossSCLR (Li et al. 2021), MCC (Su et al. 2021), AimCLR (Guo et al. 2022), ISC (Thoker et al. 2021), SkeAttnCLR (Hua et al. 2023), ActCLR (Lin, Zhang, and Liu 2023), HiCo-Transformer (Dong et al. 2023). To evaluate the merits of the proposed SCD-Net, we construct multiple downstream tasks, including action recognition, action retrieval, transfer learning and semi-supervised learning.

### Action Recognition

Here we adopt the linear evaluation method, which involves fixing the pre-trained parameters and training only a fully connected layer for label prediction. Table 1 presents a comparison of our approach with other SOTA methods on several popular datasets. The results demonstrate that our method outperforms all the existing approaches by a large margin. Specifically, we achieve 5.5% and 3.1% improvements over the previous best method on NTU-60 x-sub and x-view, respectively. On NTU-120, our approach surpasses the previous SOTA by 4.1% and 5.6% on x-sub and x-set, respectively. Again, SCD-Net achieves 91.9% on PKU-MMD I and 54.0% on PKU-MMD II, which are much higher than the existing SOTA results.

### Action Retrieval

Referring to (Thoker et al. 2021), we use the KNeighbors classifier (Cover and Hart 1967) for action retrieval while keeping all the pre-trained parameters fixed. As reported in Table 2, our SCD-Net achieves 91.9% on PKU-MMD I and 54.0% on PKU-MMD II, which are much higher than the existing SOTA results.

<table>
<thead>
<tr>
<th>Method</th>
<th>NTU-60 x-sub</th>
<th>NTU-60 x-view</th>
<th>NTU-120 x-sub</th>
<th>NTU-120 x-setup</th>
</tr>
</thead>
<tbody>
<tr>
<td>LongT GAN</td>
<td>39.1</td>
<td>48.1</td>
<td>31.5</td>
<td>35.5</td>
</tr>
<tr>
<td>P&amp;C</td>
<td>50.7</td>
<td>76.3</td>
<td>39.5</td>
<td>41.8</td>
</tr>
<tr>
<td>AimCLR</td>
<td>62.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ISC</td>
<td>62.5</td>
<td>82.6</td>
<td>50.6</td>
<td>52.3</td>
</tr>
<tr>
<td>SkeAttnCLR</td>
<td>69.4</td>
<td>76.8</td>
<td>46.7</td>
<td>58.0</td>
</tr>
<tr>
<td>HiCo-Transformer</td>
<td>68.3</td>
<td>84.8</td>
<td>56.6</td>
<td>59.1</td>
</tr>
<tr>
<td>SCD-Net (Ours)</td>
<td><strong>86.6</strong></td>
<td><strong>91.7</strong></td>
<td><strong>76.9</strong></td>
<td><strong>80.1</strong></td>
</tr>
</tbody>
</table>

Table 2: A comparison with the mainstream methods in action retrieval.
Transfer Learning For transfer learning, follow (Dong et al. 2023), we apply the knowledge representation learned from one domain to another domain. Specifically, we load the pre-trained parameters from the PKU-MMD I and NTU-60 datasets respectively, and fine-tune the model on the PKU-MMD II dataset, following the cross-subject evaluation protocol. The results presented in Table 3 demonstrate that our SCD-Net brings a performance improvement of 9.3% and 11.2%, as compared with the current SOTA results.

Semi-supervised Learning For semi-supervised learning, we first load the pre-trained parameters and then fine-tune the entire network on a partially labelled training set. In our experiment, we randomly select limited of labelled samples from the NTU-60 dataset for further training. The results in Table 4 show that even when only 1% of the labels are available, our method achieves the accuracy of 69.1% and 66.8% on x-sub and x-view, respectively. With 10% of the labelled data available, the performance of our model is further improved to 82.2% and 85.8%.

4.3 Ablation Study
In this part, we verify all the innovative components of the proposed SCD-Net. All the experimental results are focused on the action recognition task using cross-subject evaluation on the NTU-60 dataset.

The Decoupling Encoder The primary role of our novel encoder is to extract crucial spatial and temporal representations. In Table 5, when we discard the feature extractor,

the performance drops a lot. This shows that the way of extracting completely isolated information flow is not feasible in the current task, which is also in line with our expectation. It is worth noting that using non-shared feature extractors for the two branches leads to better performance than using a shared one. When we attempt to discard decouple module, compared with the default setting, the accuracy is decreased from 86.6% to 63.7% as the output is impacted by the spatiotemporal entanglement. This situation improves after converting spatiotemporal representations into temporal and spatial domain-specific embeddings, resulting in an accuracy of 84.0%. However, it was still inferior to the design with the refinement model. This is because refinement provides powerful sequence modelling capabilities, thereby refining the current rough representations.

Encoder Parameters In Table 6, we investigate the impact of the parameter settings on the model performance. Overall, the optimal performance is achieved when we use a 3-layer GCN block, 64 as the number of output channels, and set the transformer with 1 layer, 8 heads, and 2048 output channels. The results also demonstrate that changing the parameters does not significantly affect the model’s performance, indicating the stability of our approach. Additionally, we can see that the network size does not necessarily improve the performance, suggesting that it is not dependent on the network size.

Loss Function We report the results of different configurations of the loss function in Table 7. We can see that the interactive loss performs better than the traditional instance loss, leading to 0.7% and 1.6% performance boost. When using all three granularities jointly, the model achieves optimal performance. This is because there is a significant gap between the nature of the video information conveyed by
the spatial and temporal features, although they describe the same action. The global anchor provides more comprehensive representations, which bridge this gap and enhances the discriminative ability. It is worth noting that the use of both loss functions jointly does not improve the performance further. This could be attributed to the fact that the supervisory information across the information flow already provides adequate guidance, and further guidance mechanisms are unnecessary.

**Data Augmentation** Here we investigate the impact of different data augmentation strategies on the model performance. The results are reported in Table 8. Without any augmentation, the performance drops by more than 16%, compared to the default setting. When using only the conventional augmentation methods, including rotation, flipping, and shearing, the model achieves an accuracy of 85.4%. After introducing the proposed structurally guided spatiotemporal augmentation, the performance of the model increases 1.2% further. Even with random masking, the performance is still lower than the default setting.

It is worth noting that discarding either spatial or temporal masking leads to a performance degradation. Also, when only masking is used, the performance of the model is mediocre, even far worse than using only the conventional data augmentation methods. That is because our method performs a compensation, instead of replacement. A proper masking further improves the diversity of the input data and promotes the model in learning more robust spatiotemporal context associations. When combining all these techniques, the performance of the model is the best.

**Visualisation of The Decoupled Clues** As shown in Figure 4, we use t-SNE (Van der Maaten and Hinton 2008) to analyse the decoupled clues from SCD-Net. We select three groups of data with different emphases for comparison. The first row represents the spatial clue and the second one is the temporal clue. We can notice that from (a) and (d), ‘throw’ and ‘clapping’ have great separability on spatially and temporally. From (b) and (c), ‘brush teeth’ vs ‘brush hair’ are more separable on spatial domain, because the most significant difference is the object. According to (c) and (f), ‘drop’ and ‘pick up’ are more separable on temporal domain, while showing certain entanglement on the spatial domain, as they are in reverse order in temporally. More importantly, the results demonstrate that our encoder successfully decouples the corresponding features, which makes the specificity between different cues correspond to the same samples.

**5 Conclusion**

In this paper, we presented a new contrastive learning framework for unsupervised skeleton-based action recognition. The key innovation is the design of spatiotemporal clue extraction mechanism. In the proposed method, we first used a spatiotemporal modelling network to encode an action sequence, followed by a decoding module for obtaining pure spatial and temporal representations. An cross-domain loss was proposed to guide the learning of discriminative representations conveyed by different representations. The training of the system was facilitated by a novel data augmentation method tailored for the proposed unsupervised learning framework. This method imposes structural constraints on action data perturbations to enhance the efficacy of contextual modelling and increase the diversity of the data. Extensive experimental results obtained on widely used benchmarking datasets demonstrated the merits of the proposed method that defines a new SOTA of the area.

**Acknowledgements**

This work was supported in part by the National Natural Science Foundation of China (62020106012, 62332008, 62106089, U1836218), and National Key Research and Development Program of China (2023YFF1105102, 2022YFA1110200, 2023YZZS0100100).
References


