

Naive Data Augmentation Might Be Toxic: Data-prior Guided Self-supervised Representation Learning for Micro-gesture Recognition

Atif Shah, Haoyu Chen and Guoying Zhao*

Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, Finland.

Abstract—Body gestures play an important role in nonverbal communication because they transmit emotional information. Recently, a specific group of gestures, so-called Micro-gestures (MGs), has drawn increasing research interests in the community, as they can be useful cues to interpret human inner feelings. In this study, we focused on recognizing MG via self-supervised learning from skeleton sequences with several contributions. Initially, we observed that existing data augmentation methods for skeleton data always fail in MG representation learning. Our investigation shows that the failure is caused by the inherent properties of real-world datasets, such as imbalanced/long-tail data distribution, intra-class ambiguity, and inter-class heterogeneity. Thus, we propose a novel prior-guided augmentation strategy that can preserve the original data distribution while maximizing the agreement between samples in self-supervised learning. Furthermore, we proposed a three-stream architecture of self-supervised presentation learning for micro-gestures via spatial/temporal masking to jointly enhance the learning of invariant features. Lastly, the experimental results show that our proposed method has achieved state-of-the-art performances on two public MG datasets.

I. INTRODUCTION

Human body gestures play a vital role in non-verbal communication for transmitting a wide range of emotional information to others in social contexts [1], [2]. However, current computer emotional cue analyses usually facilitate vocal and facial emotions, while overlooking the essential emotional nuances communicated by body gestures [3], [4].

Recently, a specific group of gestures, so-called Micro-gestures (MGs), has drawn increasing research interests in the community, as they have a significant advantage over other modalities for human emotion understanding[5], [6]. Firstly, obtaining data on body gestures is more feasible, especially in public areas where high-resolution surveillance cameras for facial expressions or microphones for voice capturing are not widely available. Secondly, recent advances in deep learning on large-scale datasets have sparked concerns about human privacy [7]. As a result, body gestures have a higher degree of privacy for person identification. Another advantage is ordinary people can hide their emotions using false facial expressions due to the social norm; however, only a few could manage to hide their true emotions via body gestures [8]. Thus, the MG, a group of unintentional behaviors, like rubbing hands under stress, which is caused by one’s inner feelings, has drawn researchers’ attention in recent years. These are largely uncontrollably emotional gestures that reveal an individual’s actual feelings. MG are

* indicates corresponding author

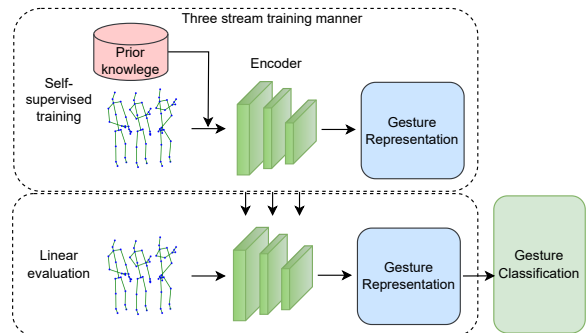


Fig. 1. A generic self-supervised learning framework. We first use data prior-guided data augmentation method to enhance the self-supervised learning of the gesture presentation (top). Then the learned presentation will be utilized to conduct linear evaluation.

more difficult to recognize than typical gestures and actions since they are quick, delicate, and used to conceal a person’s inner feelings [9], [10]. However, the inherent properties of MG datasets limit the direct implementation of data-driven learning approaches in related tasks [11]. For instance, a great deal of annotated data is necessary for existing mainstream deep-learning methods to learn strong visual representations effectively, while the available labeled MGs are limited by severely imbalanced/long-tail data distribution problems. Towards this data imbalance issue, there has been progress in exploring self-supervised learning methods [10], which help reduce the reliance on labeled data and make models learn visual representations that are similar to those achieved through traditional supervised learning, especially in areas like image categorization [12]. However, when it comes to action, gestures, and micro-gesture (MG) recognition, using a self-supervised method hasn’t progressed much, and the best way to apply a self-supervised method in this context is still an unanswered question.

With the above observations, we study how to enhance MG representation learning in a self-supervised manner with several key contributions. The initial contribution is that we explicitly show that existing data augmentation strategies for skeleton data will dramatically decrease the performance of the self-supervised learning, as naive data augmentation [13] will not only damage the original real-world long-tail distribution but also introduce noise in which semantic information is destroyed. Thus, we propose a novel prior-guided data augmentation method for self-supervised learning that can comply better with the natural MG data distributions.

Besides, we propose a novel three-stream self-supervised architecture for MG recognition. The first two streams use a spatial/temporal masking strategy to mask the joints/frames with a high level of centrality. This enables the network to learn the robust relationship between joints/frames. The third anchor stream passes the original input sequence with conventional augmentation to keep the semantic information of gestures for both the spatial and temporal streams for cross-correlation. Figure 1 illustrates the generic framework of our method for self-supervised MG recognition, where the self-supervised learning step uses prior-guided knowledge to learn more robust MG representations that are utilized during the linear evaluation step for MG classification.

The primary contributions in this work are as follows:

- We propose a self-supervised micro-gestures representation learning (MGRL) framework with three streams to enhance spatial and temporal learning.
- We investigate how the existing data augmentation approaches will fail on the real-world skeleton data and demonstrate the causes.
- We further introduce a novel skeleton-merge data augmentation strategy guided by the prior learned from the data distribution to improve the performance and robustness of the framework.
- The efficiently designed framework achieves state-of-the-art results on two MG datasets.

II. RELATED WORK

Gestures are a particular kind of action that frequently consists of intentional, expressive signals or bodily motions and is commonly used in nonverbal communication to convey information or express emotions. Therefore, we consider action recognition methods in this work.

A. Supervised skeleton-based action recognition

Skeleton-based data consist of 2D and 3D coordinates of human body joints and have advantages over videos and imaging data [14]. Skeleton data are more resistant to disruptions, such as illumination changes and background clutter, allowing for a more effective analysis of gestures and actions [14]. Spatial-Temporal Graph Convolutional Network (STGCN) [15] extracts detailed information from human skeletons to improve action recognition. Si et al. [16] combined LSTM with convolutional networks to improve overall performance by leveraging discriminatory spatial and temporal information. For action recognition, Shi et al. [17] used a Two-Stream Adaptive Graph Convolutional Network with a data-dependent graph with first and second-order bone information. Unlike STGCN, which connects a single node to its neighbors in temporal space, Liu et al. [18] introduced the Multiscale Unified Spatial-Temporal Graph Convolutional Operator (MSG3D), which introduces a dense connection to the next temporal nodes to extract more detailed features. Despite these advances, the aforementioned approaches require a large amount of annotated data to acquire the feature representation of the target data distribution.

B. Self-supervised skeleton-based action recognition

Existing studies on self-supervised representation learning investigate numerous pretext tasks to capture motion as context. LongTGAN [19] learned 3D action representation using sequence reconstruction. P&C [20] improved learning representation using weak decoders. Motion prediction and jigsaw puzzle tasks used by MS2L [21] and Yang et al. [22] proposed skeleton cloud colorization task. Contrastive learning approaches gained popularity in 3D action recognition, such as AS-CAL [23] and SkeletonCLR [24], make use of momentum encoders and provide various data augmentation strategies. ActCLR [25] focused on adaptive action modeling for different body parts, whereas AimCLR [26] introduced extreme augmentation. Despite being effective, contrastive learning algorithms frequently overlook important local spatio-temporal information required for 3D action modeling. The increasing popularity of transformers has encouraged self-supervised pretraining for visual representation learning based on masked visual modeling [27], [28]. Masked Autoencoder (MAE) approaches are used in SkeletonMAE [13] and MAMP [29] to learn 3D action representations. SkeletonMAE reconstructs spatial coordinates using a skeleton-based encoder-decoder transformer, whereas MAMP employs Masked Motion Prediction to explicitly characterize temporal action.

The proposed method uses a three-stream architecture inspired by the Barlow Twins. This upper stream uses spatial masking and the lower stream uses temporal masking, as well as an additional anchor stream in the middle to keep the original semantic information.

To elaborate further, the proposed framework is a foundational negative-sample-free three-stream network that uses conventional augmentation and Skeleton-merge augmentation for self-supervised skeleton learning.

The upper stream first modifies the computing mode of the GCN backbone to ensure that information from masked joints is removed from the feature computation process. As a result, joints with a higher degree of centrality are more likely to be hidden.

The lower stream computes the motion value corresponding to each frame of the gesture, which is used as the attention weight for picking masked frames.

III. METHODOLOGY

We presented a self-supervised architecture for MG recognition that uses a three-stream method as shown in Figure 2. In Figure 2, for the upper stream, a spatial masking method is used to obscure joints with a high level of centrality. This enables the network to determine the relationship between masked and unmasked joints during cross-correlation with the anchor stream. Meanwhile, the lower stream uses frame masking to focus on frames with a large number of gesture moments. This illustrates the interplay of redundant and important frames. The anchor stream keeps the original input sequence while employing standard augmentation to retain gesture semantics. This method guarantees cross-correlation between the upper and lower streams. A linear classifier

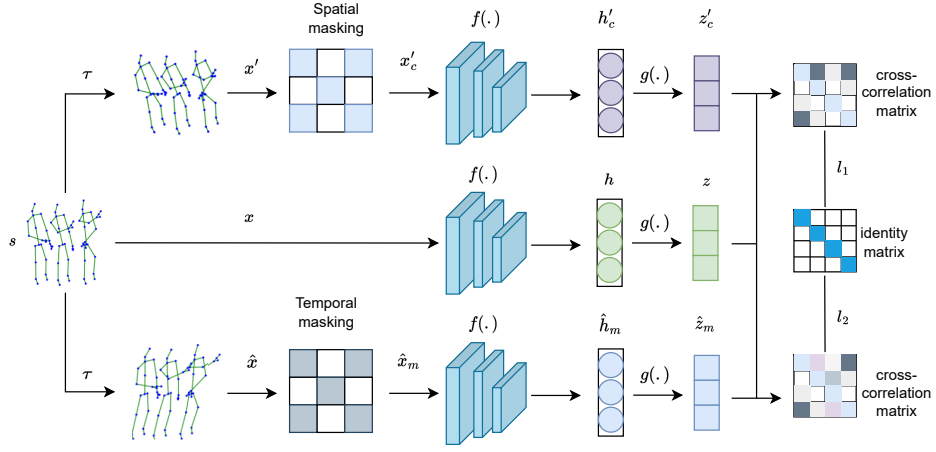


Fig. 2. A self-supervised architecture for MG recognitions. We introduce three streams to enhance self-supervised learning through spatial and temporal masking.

is added at the top of a trained model to achieve MG recognition.

A. Self-supervised framework

We first introduce our proposed self-supervised framework to obtain a representation of skeleton-based gestures, taking inspiration from the recent advances in negative-sample-free self-supervised methods [30], [31]. These methods do not require negative samples, nor do they require a large batch size or memory bank. This involves using a shared encoder in a symmetrical network to generate features.

We represent a 3D human skeleton sequence input as $s \in \mathbb{R}^{C \times T \times V}$, where C denotes the 3D skeleton’s channel dimension, and T represents frames and V joints.

The conventional augmentation τ applied on input skeleton sequence s to extract the diverse views x' , x and \hat{x} and passed it to the encoders. The three encoders f' , f and \hat{f} extract features $h' = f(x')$, $h = f(x)$ and $\hat{h} = f(\hat{x})$, respectively where h , h' and $\hat{h} \in \mathbb{R}^{C_h}$. The self-attention layer processes the features and extracts intrinsic features. Subsequently, each feature is mapped onto a higher-dimensional space using a projector g to get the embeddings $z' = g(h')$, $z = g(h)$ and $\hat{z} = g(\hat{h})$, where z' , z and $\hat{z} \in \mathbb{R}^{C_z}$. Finally, the encoder can successfully capture the relationship between the two streams and the anchor stream by encouraging the empirical cross-correlation matrix between the embeddings of distorted variations to create an identity matrix.

B. Spatial and Temporal masking

To enhance the joint resilience of the learned representation, this module favors the partial skeleton data to develop

features similar to the entire skeleton data. The encoder records the interaction between masked and unmasked joints, which is critical for the challenging downstream task. It is important to build a partial skeleton data for the encoder.

Previous study shows that changing the masked joint values to explicitly zero out is unreasonable as the joints involved semantic information involving 3D joint location that varies. Therefore the STGCN computation mode is modified and the masked skeleton joints are excluded from the corresponding rows and columns in the adjacency matrix. This process will exempt the masked joints from processing and involve only unmasked joints to generate features.

We take advantage of the human skeleton graph topology’s degree centrality, noticing that joints with greater degrees can gather more extensive neighborhood information. As a result, we assign a larger probability to mask such connected joints which will allow the encoder to capture relationships across a broader range. Masking such a centralized joint allows the encoder to learn the relationship among a large number of surrounding joints. The degree d_i is calculated for each skeleton joint V_i , where $V_i, i \in (1, 2, \dots, n)$, where n indicates the total number of joints, and sets their masked probability:

$$p_i = \frac{d_i}{\sum_{j=1}^n d_j} \quad (1)$$

In temporal masking, temporal information can be more redundant and semantic information is concentrated as compared to the spatial joints. As most of the frames have little information while certain key-frames convey some curial semantic gesture information. To further elaborate, we seek to locate these key frames and use masking to produce a difficult pair for the anchor stream and temporal

masking stream within our proposed method. The encoder can effectively capture the relationship between redundant frames and key-frames by encouraging the empirical cross-correlation matrix between the embeddings of two streams to be an identity matrix. In particular, the motion denoted as $m \in \mathbb{R}^{C \times T \times V}$ for the sequence s is computed using the time interval between frames $m_t = x_{t+1} - x_t$. Next, we obtained the motion rate of a frame that acts as an attention-weight.

$$a_i = \frac{m_t^2}{\sum_{i=1}^T m_i^2} \quad (2)$$

The top-K attention weights a_{i_1}, \dots, a_{i_K} are chosen and the sequence x_{t_1}, \dots, x_{t_K} serves as key-frames that carry additional semantic information about the gesture. We can find some frames with high motion in the middle of the sequence and these are the key-frames. We masked these frames and encouraged the features from the masked sequence to be close to the anchor feature that encloses the entire semantic information

C. Objective function

The conventional augmentation is applied to obtain views x, x', \hat{x} from the input skeleton sequence s . The upper stream uses x' view and applies spatial masking to generate a partial skeleton data x'_c . Similarly, the lower stream applies temporal masking on \hat{x} view and obtains \hat{x}_m . To keep the original semantic information, the middle anchor stream uses x view and do not use any masking strategy. To extract the features, a shared encoder is used which operates simultaneously. The learned features are denoted as $h = f(x), h' = f(x'), \hat{h}_m = f(\hat{x}_c)$ with learned parameter θ each. The embeddings are extracted via project g in high dimensional space, which are denoted as $z = g(h), z' = g(h')$ and $\hat{z} = g(\hat{h})$. The cross-correlation matrix is used to represent the relationship between masked and unmasked joints, denoted as C' between z and z'_c embeddings. Following the C' calculation, the l_1 loss is applied.

$$C'_{ij} = \frac{\sum_b z_{b,i} z'_{b,j}}{\sqrt{\sum_b (z_{b,i})^2} \sqrt{\sum_b (z'_{b,j})^2}} \quad (3)$$

where b represents batch dimension and i, j show the embedding dimension.

$$l_1 = \sum_i (1 - C'_{ii})^2 + \lambda \sum_i \sum_{j \neq i} C'_{ij}{}^2 \quad (4)$$

where the initial term stimulates the diagonal element of C' to 1. This ensures that partial data is represented similarly to the entire data. The following term decouples the embedding components, reducing redundancy and preventing

the representation from becoming constant. The λ keeps the dimension difference balanced between both terms.

Likewise, l_2 loss is calculated for z and \hat{z}_m

$$l_2 = \sum_i (1 - \hat{C}_{ii})^2 + \lambda \sum_i \sum_{j \neq i} \hat{C}_{ij}{}^2 \quad (5)$$

where \hat{C} represents the cross-correlation matrix between z and \hat{z} . The l_2 loss characterizes the connection between masked and unmasked frames. Finally, the total loss is calculated as

$$l = l_1 + l_2 \quad (6)$$

IV. A NOVEL PRIOR-GUIDED SKELETON-MERGE AUGMENTATION

We first investigate the drawbacks of the existing data augmentation methods and then propose our newly designed data augmentation strategy by addressing the issues existing in the current methods.

A. Investigation of existing data augmentation methods

Aside from common data augmentation operations like sheering, rotation, and cropping (see also in the experimental section), there are also advanced data augmentation strategies customized for skeleton data such as SkeletonMix [32]. The mixup augmentation was initially designed for image data augmentation [33], [32]. The two images are combined with their labels linearly by

$$\begin{aligned} \bar{x} &= \mu x_i + (1 - \mu) x_j \\ \bar{y} &= \mu y_i + (1 - \mu) y_j \end{aligned} \quad (7)$$

where image represented as (x_i, y_i) , label as (x_j, y_j) and output sample as (\bar{x}, \bar{y}) . The bias between two input samples is controlled by using $\mu \in [0, 1]$. We can not combine the skeleton directly like images, because of deformation, which leads to poor performance.

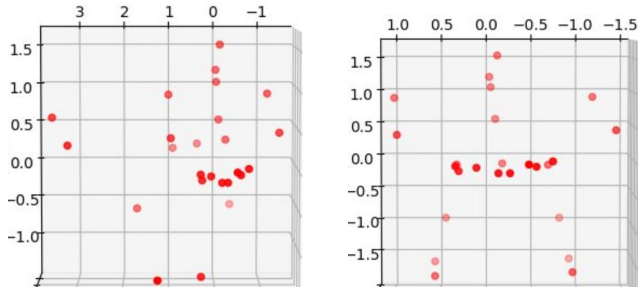
We used the above strategy to merge the upper and lower body skeletons of two randomly selected individuals.

$$\begin{aligned} \bar{x} &= Merge(x_i, \mathcal{J}_u, x_j, \mathcal{J}_l) \\ \bar{y} &= \mu y_i + (1 - \mu) y_j \end{aligned} \quad (8)$$

where \mathcal{J}_u and \mathcal{J}_l represent the upper and lower body joints, respectively.

However, we discovered that explicitly joining and merging two skeletons did not produce adequate results, as described in the Experiment section. In Table II, we can see that naively incorporating the Skeleton-merge data augmentation into the existing framework will lead to a dramatic decrease in the performances for more than 20%.

By further investigating the causes of the failure, we find that two main sources are harmful to learning. Firstly, by



(a) Existing data-augmentation method Skeleton-mix. (b) After normalization and instance-wise centralization (ours).

Fig. 3. A visualized demonstration showing the existing Skeleton-mix [32] method. We can see that directly implementing this data augmentation will lead to noisy and harmful samples in which semantic information is destroyed.

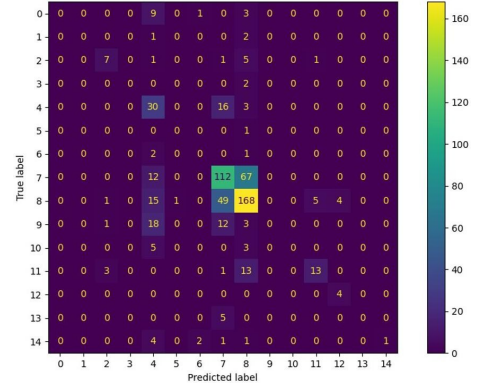
TABLE I

AUGMENTATION RATIO ON SMG DATASET SIZE. * MEANS IT EQUALS TO NO DATA AUGMENTATION.

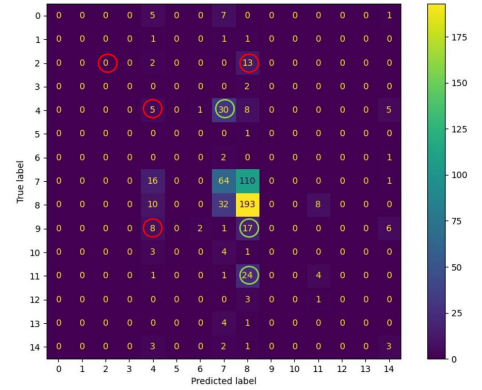
Dataset	% augmentation ratio	Accuracy
SMG	0%*	52.18
	10%	58.45
	20%	59.06
	30%	57.65
	100%	38.28

visualizing the data augmentation results as shown in Figure 3, we can see that directly implementing Skeleton-mix [32] data augmentation will lead to noisy and harmful samples in which semantic information is destroyed. Besides, in Table I, we also find that the dramatic performance decrease caused by the data augmentation is strongly associated with the size of the ratio. When the dataset is augmented up to double size (the last line in Table I), the performance drops more than 20% compared to the non-augmented one. We assume that the data augmentation damaged the original data distribution by improperly scaling the dataset size. To confirm how the data augmentation (with improper scaling ratio) damaged the data distribution, we visualized the confusion matrix of the MG recognition results as shown in Figure 4, we can observe that the data-augmented one strengthened the bias to the specific MG categories, leading to more false positive predictions. In other words, data augmentation with improper ratio size will weaken the prior of those MG categories with fewer samples.

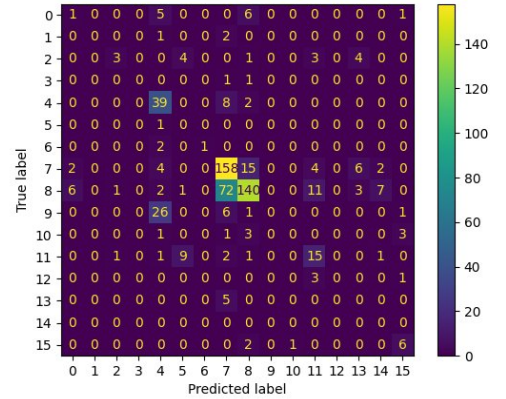
Thus, in the next section, we introduce how to use strategies that emphasize the skeleton and incorporate previous knowledge.



(a) Confusion matrix without data augmentation



(b) Confusion matrix with naive data augmentation



(c) Confusion matrix with prior-guided data augmentation (ours)

Fig. 4. Confusion matrix of different augmentation methods.

B. A novel prior-guided data augmentation for self-supervised learning

We wish to design a data augmentation technique that can precisely address scenarios in which 1) mitigate the noises brought by samples in which semantic information is missing; 2) the data-augmented dataset will have a similar data distribution as the ordinary one; and 3) certain samples underperformed due to similarities in skeleton structure, resulting in model confusion. We achieve them via the following steps:

Normalization and centralization. We conduct normalization firstly to the whole dataset to ensure all the samples that going to be merged are aligned in the same feature space. Then, we conduct instance-wise centralization before merging two samples, leading to a more natural combination of the two samples. Specifically, we use one sample’s coordinates as the anchor coordinate system, then merge the other samples to this sample within the same coordinate system via translation. The resulting Skeleton-merge samples are shown in Figure 5 and Figure 6. Figure 5 shows two different skeletons from “rubbing hands” and “moving legs” labels, after performing skeleton-merge both the skeletons are combined into one skeleton as illustrated in Figure 6, which contains more meaningful semantic information.

Euclidean distance as prior. As discussed above, samples contain ambiguity in the dataset. For example, the “Moving legs” class is commonly misidentified as “Crossing legs,” and vice versa. We use such information combined with distance maps between classes as prior guidance for the model to better understand the complex and similar skeleton structures. We deploy the Euclidean distance to build up distance maps among different samples, which is crucial to identify patterns among classes that can be used as prior guidance, as shown in Figure 7.

Prior-guided data augmentation. After we obtain the prior as shown in Figure 7, we augment the dataset based on this prior by screening the most confusing MG categories (based on the Euclidean distance, from the lowest to the highest) and conducting the Skeleton-merge on those “hard” samples. In this way, we not only preserve the prior of data distribution but also enhance the model learning on those “hard” samples with augmented variants.

V. EXPERIMENTS AND RESULTS

A. Micro-gestures datasets

We use two real-world datasets collected from in-the-wild settings. The Spontaneous Micro-Gesture (SMG) dataset [5] has 3,692 samples of 17 MG. The dataset is acquired using

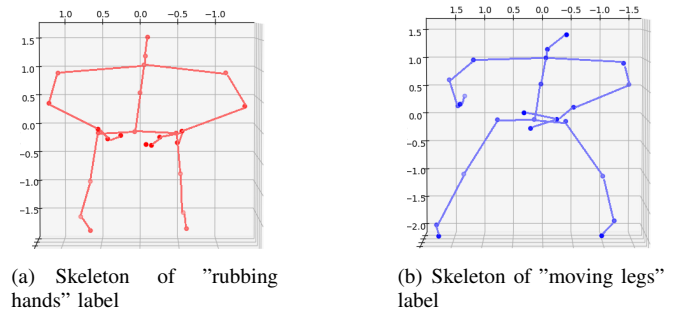


Fig. 5. Skeleton of two different classes

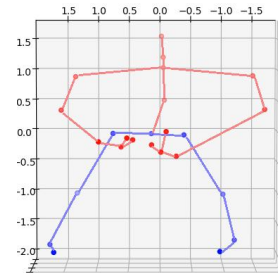


Fig. 6. Skeleton-merge of two different classes

Kinect from 40 people who narrated both fake and actual stories. The data was collected in 25 3D skeleton joints.

The Micro-Gesture Understanding and Emotion Analysis (iMiGUE) dataset [6] contains 32 MG collected from post-match press conference videos. The training set includes 13,936 samples, whereas the testing set contains 4,563 MG samples for recognizing negative and positive emotions. The data is acquired using OpenPose and accumulated 25 3D skeleton joints.

B. Implementation details

We used preprocessing methods from AimCLR [26] and restricted the frame size to 50. The STGCN [15] with 16 hidden layer configurations is used as the backbone. Adam optimizer is used for representation learning and downstream tasks with 200 epochs and mini-batch size 32. Before passing data to streams, the data augmentation is performed which includes three spatial and one temporal augmentation. we use the shear, crop, rotate, and spatial flip augmentation.

Shear is the linear transformation of the skeleton data which can be achieved using a shear matrix multiplied by 3D coordinates of the human skeleton, which will generate a random shear angle. *Cropping* is used to increase data diversity by first padding a portion of the frames in the original sequence and then conducting a random crop to restore the

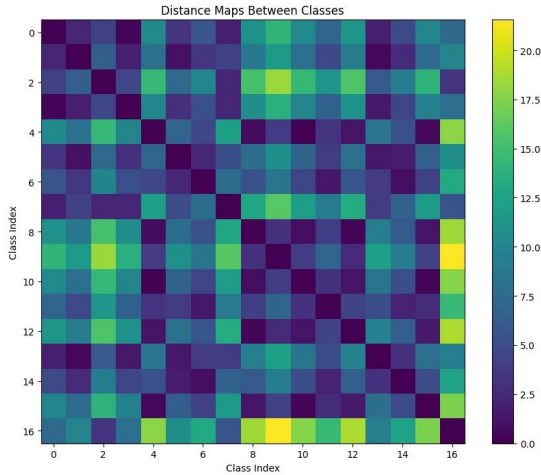


Fig. 7. Distance maps between classes

TABLE II

ABLATION STUDY OF DIFFERENT AUGMENTATION STRATEGIES ON SMG DATASET. NOTE THAT THE SKELETON-MERGE IS THE SAME AS SKELETON-MIX BUT ONLY WITH UPPER AND LOWER BODY MERGE.

Accuracy	Augmentation	Prior-guided
52.18	No data augmentation	No
38.28	Skeleton-merge	No
39.53	Skeleton-merge + flip hands	No
47.03	Skeleton-merge + centralization	No
58.45	Skeleton-merge + centralization	Yes (threshold 1)
63.50	Skeleton-merge + centralization	Yes (threshold 2)

sequence to its original length. *Rotation* is another effective technique that allows the representation to adjust to variations in spatial spaces efficiently. *Spatial flip* augmentation flips the body of the skeleton left-to-right with 50% probability at a time.

C. Ablation study

We conducted an ablation study to choose the appropriate size of the prior-guided augmentation which is shown in Table II and Table I.

Augmentation could be very challenging if we randomly merge two skeletons, results are shown in Table II. The second row shows the skeleton-merge augmentation on SMG datasets with no prior-guide one and we get 38.28% accuracy. In addition to skeleton-merge, we introduced flip hand augmentation where we flip the left and right hands of skeletons, and the results improved slightly with 39.53% accuracy as shown in the second row. One of the important aspects of skeleton-merge is that the skeleton should be centralized when we perform Skeleton-merge, as the results

TABLE III

COMPARING OUR RESULTS WITH STATE-OF-THE-ART METHODS

Dataset	Method	Accuracy
SMG	MSG3D [18]	44.6
	A-EDGCN [10]	47.9
	PSTL [31]	58.4
	MGRL(ours)	63.5
iMiGUE	P&C [20]	31.7
	U-S-VAE Z [6]	32.4
	MSG3D [18]	36.9
	A-EDGCN [10]	37.5
	PSTL [31]	37.6
	MGRL(ours)	38.5

indicate in the third row of Table II, which achieved 47.03% accuracy. Lastly, the prior guidance is added to improve the performance, which includes those classes that were under-performed previously now improved the overall performance and achieved 58.45%/63.50% accuracy (with threshold 1 and 2), as shown in the last two rows of Table II. In both thresholds, we employed the prior-guided skeleton merge augmentation for moving legs and crossing legs. We applied a 10% augmentation for threshold 1 and a 20% augmentation for threshold 2. Note that the threshold is determined to screen the “hard” categories to merge in the prior-guided data augmentation.

One of the challenging issues is how many samples should be augmented and in which ratio for each category to get the optimal performance, according to our preliminary experiment shown in Table I. The experiments are conducted with a certain percentage of the sample size as shown in the second column of Table I. The first row indicates no augmentation, and the second row shows that 10% augmentation slightly improves the performance. Similarly, with 20% augmentation, the results are improved further and achieved 59.06% accuracy as shown in the third row. A too-high augmentation ratio (100%) will lead to a dramatic decrease in performance (by more than 20%). Thus, we introduce the prior-guided data augmentation that automatically selects the “hard” categories to augment in ratios that align with the data distribution. As shown in Table II, the prior guided augmentation is much better than any compared methods.

D. Comparison with state-of-the-art methods

Table III shows the comparison results with previous methods on two MG datasets, SMG and iMiGUE where the bold text represents the highest accuracy. Our proposed method outperformed some supervised methods such as MSG3D and unsupervised and self-supervised methods.

To elaborate further, augmentation is a challenging issue to tackle when dealing with real-world skeleton data. Combining two different skeletons explicitly would not perform as desired, and could lead to skeleton deformation if not handled carefully, as shown in Figure 3. The main reason is that the object could be moving back and forth among the frames which could lead to skeleton deformation during skeleton-merge. Before each skeleton-merge augmentation, the skeleton should be normalized and centralized as shown in Figure 3a. This way, the skeleton maintains its shape, while on the other hand, the non-centralized skeleton faces severe deformation, which decreases the performance and accuracy.

Similarly, the overall accuracy and class-wise accuracy also drop significantly as shown in Figure 4. This figure shows the confusion matrix for both original skeletons and augmented skeletons results, where the skeleton-merge augmentation is performed without normalization and centralization. The number of correctly classified samples dropped significantly. For instance, in the confusion matrix of SMG data, the label 4 results dropped from 30 to 5 and the misclassification increased from 16 to 30 samples. Similarly, the green and red circles show the increase and decrease in the sample size between SMG data and augmented SMG data, respectively.

VI. CONCLUSIONS

In this work, we proposed a self-supervised three-stream method with spatial and temporal masking. The anchor stream preserves semantic information, while the encoders establish a connection between the three-streams and the anchor stream. This is achieved by promoting the empirical cross-correlation matrix between the embeddings of distorted variations to form an identity matrix. We proposed and investigated the prior-guided skeleton-merge augmentation strategy for real-world skeleton datasets. The experiments were conducted on two micro-gesture datasets, SMG and iMiGUE, and achieved state-of-the-art results of 63.5% and 38.5%, respectively.

VII. ACKNOWLEDGEMENTS

This work was supported by the Research Council of Finland (former Academy of Finland) Academy Professor project EmotionAI (grants 336116, 345122), the University of Oulu & Research Council of Finland Profi 7 (grant 352788). The authors also wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

- [1] H. Aviezer, Y. Trope, and A. Todorov, “Body cues, not facial expressions, discriminate between intense positive and negative emotions,” *Science*, vol. 338, no. 6111, pp. 1225–1229, 2012.
- [2] G. Zhao, Y. Li, and Q. Xu, “From emotion ai to cognitive ai,” *International Journal of Network Dynamics and Intelligence*, pp. 65–72, 2022.
- [3] S. Li and W. Deng, “Deep facial expression recognition: A survey,” *IEEE transactions on affective computing*, vol. 13, no. 3, pp. 1195–1215, 2020.
- [4] H. Shi, W. Peng, H. Chen, X. Liu, and G. Zhao, “Multiscale 3d-shift graph convolution network for emotion recognition from human actions,” *IEEE Intelligent Systems*, vol. 37, no. 4, pp. 103–110, 2022.
- [5] H. Chen, H. Shi, X. Liu, X. Li, and G. Zhao, “Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis,” *International Journal of Computer Vision*, vol. 131, no. 6, pp. 1346–1366, 2023.
- [6] X. Liu, H. Shi, H. Chen, Z. Yu, X. Li, and G. Zhao, “imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 631–10 642.
- [7] S. J. Oh, R. Benenson, M. Fritz, and B. Schiele, “Faceless person recognition: Privacy implications in social media,” in *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 19–35.
- [8] P. Ekman, “Darwin, deception, and facial expression,” *Annals of the new York Academy of sciences*, vol. 1000, no. 1, pp. 205–221, 2003.
- [9] H. Chen, X. Liu, X. Li, H. Shi, and G. Zhao, “Analyze spontaneous gestures for emotional stress state recognition: A micro-gesture dataset and analysis with deep learning,” in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–8.
- [10] A. Shah, H. Chen, H. Shi, and G. Zhao, “Efficient dense-graph convolutional network with inductive prior augmentations for unsupervised micro-gesture recognition,” in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 2686–2692.
- [11] H. Chen, H. Tang, Z. Yu, N. Sebe, and G. Zhao, “Geometry-contrastive transformer for generalized 3d pose transfer,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 258–266.
- [12] X. Chen, H. Fan, R. Girshick, and K. He, “Improved baselines with momentum contrastive learning,” *arXiv preprint arXiv:2003.04297*, 2020.
- [13] W. Wu, Y. Hua, C. Zheng, S. Wu, C. Chen, and A. Lu, “Skeletonmae: Spatial-temporal masked autoencoders for self-supervised skeleton action recognition,” in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE, 2023, pp. 224–229.
- [14] L. L. Presti and M. La Cascia, “3d skeleton-based human action classification: A survey,” *Pattern Recognition*, vol. 53, pp. 130–147, 2016.
- [15] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.
- [16] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, “An attention enhanced graph convolutional lstm network for skeleton-based action recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1227–1236.
- [17] L. Shi, Y. Zhang, J. Cheng, and H. Lu, “Two-stream adaptive graph convolutional networks for skeleton-based action recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 026–12 035.
- [18] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, “Disentangling and unifying graph convolutions for skeleton-based action recogni-

- tion,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 143–152.
- [19] N. Zheng, J. Wen, R. Liu, L. Long, J. Dai, and Z. Gong, “Unsupervised representation learning with long-term dynamics for skeleton based action recognition,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
- [20] K. Su, X. Liu, and E. Shlizerman, “Predict & cluster: Unsupervised skeleton based action recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9631–9640.
- [21] L. Lin, S. Song, W. Yang, and J. Liu, “Ms2l: Multi-task self-supervised learning for skeleton based action recognition,” in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2490–2498.
- [22] S. Yang, J. Liu, S. Lu, M. H. Er, and A. C. Kot, “Skeleton cloud colorization for unsupervised 3d action representation learning,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 423–13 433.
- [23] H. Rao, S. Xu, X. Hu, J. Cheng, and B. Hu, “Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition,” *Information Sciences*, vol. 569, pp. 90–109, 2021.
- [24] L. Li, M. Wang, B. Ni, H. Wang, J. Yang, and W. Zhang, “3d human action representation learning via cross-view consistency pursuit,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4741–4750.
- [25] L. Lin, J. Zhang, and J. Liu, “Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2363–2372.
- [26] T. Guo, H. Liu, Z. Chen, M. Liu, T. Wang, and R. Ding, “Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 762–770.
- [27] H. Bao, L. Dong, S. Piao, and F. Wei, “Beit: Bert pre-training of image transformers,” *arXiv preprint arXiv:2106.08254*, 2021.
- [28] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.
- [29] Y. Mao, J. Deng, W. Zhou, Y. Fang, W. Ouyang, and H. Li, “Masked motion predictors are strong 3d action representation learners,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 181–10 191.
- [30] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, “Barlow twins: Self-supervised learning via redundancy reduction,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 310–12 320.
- [31] Y. Zhou, H. Duan, A. Rao, B. Su, and J. Wang, “Self-supervised action representation learning from partial spatio-temporal skeleton sequences,” *arXiv preprint arXiv:2302.09018*, 2023.
- [32] K. Xu, F. Ye, Q. Zhong, and D. Xie, “Topology-aware convolutional neural network for efficient skeleton-based action recognition,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2866–2874.
- [33] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” *arXiv preprint arXiv:1710.09412*, 2017.