Skeleton-based Self-Supervised Feature Extraction for Improved Dynamic Hand Gesture Recognition

Omar Ikne, Benjamin Allaert, Hazem Wannous

2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Skeleton-based Self-Supervised Feature Extraction for Improved Dynamic Hand Gesture Recognition

Omar Ikne[1], Benjamin Allaert[1] and Hazem Wannous[1]
[1] IMT Nord Europe, Institut Mines-Télécom, Univ. Lille, Centre for Digital Systems, F-59000 Lille, France

***Abstract*— Human-computer interaction (HCI) has become integral to modern life, especially in digital environments. However, challenges persist in utilizing hand gestures due to factors such as the dynamic nature of gestures and the intricacies of intra and inter-finger movements. In this paper, we propose an innovative approach to improve skeleton-based hand gesture recognition by integrating self-supervised learning, a promising technique for acquiring distinctive representations directly from unlabeled data. The proposed method takes advantage of prior knowledge of hand topology, combining topology-aware self-supervised learning with a customized skeleton-based architecture to derive meaningful representations from skeleton data under different hand poses. We introduce customized masking strategies for skeletal hand data and design a model architecture that incorporates spatial connectivity information, improving the model's understanding of the interrelationships between hand joints. The extensive experiments demonstrate the effectiveness of the approach, with state-of-the-art performance on benchmark datasets. An exploration of the generalization of learned representations across datasets and a study of the impact of fine-tuning with limited labeled data are conducted, highlighting the adaptability and robustness of our approach. Codes are available at: `https://github.com/o-ikne/SkelMAE`.**

I. Introduction

Dynamic Hand Gesture Recognition (HGR) is an essential component of human-computer interaction systems, facilitating natural, intuitive communication between human and machine. Its applications cover a wide range of fields, from sign language interpretation to virtual reality control and beyond. In recent years, advances in deep learning and computer vision techniques have enabled significant advances in hand gesture recognition systems [31].

Traditional approaches for hand gesture recognition often rely on color or depth image data. However, these methods can be limited in capturing fine hand movements and the complex spatial relationships between hand joints [37]. Meanwhile, skeleton-based representations, which are extracted from RGB and depth images, have gained in popularity. They provide a structured and informative way of representing hand poses, reducing computational complexity and privacy concerns, and offering potential improvements in recognition accuracy[49]. Moreover, unlike traditional RGB-based approaches, skeleton-based representations are more robust to variations in lighting, camera viewpoints and other background changes [39].

Recent advances on dynamic hand gesture recognition rely on fully-supervised learning methods on 3D skeleton coordinates using Convolutional Neural Networks (CNNs) [47], Convolution Graph Networks (CGNs) [38] and Transformers [12]. However, fully supervised learning methods can be prone to overfitting and require a large amount of labeled training data. To address these issues, self-supervised learning approaches, have been increasingly frequent [16], [45], particularly with the emergence of Masked Autoencoders (MAEs) [16] based on Vision Transformer (ViT) [10]. Nevertheless, ViT was originally designed for image processing, posing a challenge when adapting it to skeleton-based data. The incorporation of a self-supervised learning phase based on MAEs is a promising approach that has emerged [45]. The self-supervised phase enables MAEs to autonomously extract discriminative features and representations from raw skeleton data, without the need for manual labeling. The resulting learned representations form a valuable basis for the training of later classification models, enhancing their performance in gesture recognition tasks. Nevertheless, while self-supervised skeletal learning has found extensive application in human action recognition [45], [19], its use in the context of hand gesture recognition remains relatively unexplored.

In this paper, we presents an innovative approach to improve skeleton-based hand gesture recognition by leveraging the strength of self-supervised learning. Self-supervised learning has emerged as a promising paradigm for pre-training models on unlabeled data, enabling them to capture rich, contextual features. By adapting self-supervised learning techniques to skeleton data, our objective is to learn meaningful representations of hand gestures that generalize across different hand poses and gestures.

Inspired by MAEs [16], we propose a masked autoencoder for self-supervised skeleton learning. The proposed approach introduces several key innovations. First, we explore various masking strategies for skeleton self-supervised pre-training, enabling the model to learn the importance of different joints by partially masking subsets of skeleton joints. Secondly, we adapt the ViT architecture to process skeleton-based data and incorporate spatial connectivity information using adjacency matrices, enhancing the model's understanding of spatial relationships between hand joints. Finally we use Spatial Temporal Graph Convolutional Network (STGCN) [46] as a backbone for dynamic hand gesture classification based on the newly learned representations.

Throughout our experiments, we demonstrate that the proposed approach achieves state-of-the-art results on benchmark datasets, including Briareo [29], IPN Hand [4] and SHREC'17 dataset [8]. We evaluate the model's performance in both single and cross-dataset settings, highlighting its adaptability and robustness. Furthermore, we explore fine-

979-8-3503-9494-8/24/$31.00

tuning with limited labeled data, showcasing the model's efficient generalization.

In summary, our contributions in this work include:

- The Proposal of an efficient MAE for the hand skeleton, adapting the ViT architecture for skeletal data processing. The adapted ViT has some novelties including:
  - Integration of Fourier feature mapping, showcasing its superiority over linear mapping in capturing spatial relationships.
  - A modified attention mechanism formula that incorporates adjacency matrix information, enhancing joints spatial connectivity encoding.
- The implementation, and evaluation of various masking strategies tailored for hand skeleton data to facilitate distinctive representation learning.
- The evaluation of our method on the Briareo, IPN Hand and SHREC'17 datasets, and demonstration of its state-of-the-art performance.

The paper is structured as follows: First, in Section II, we review the relevant literature on hand gesture recognition and self-supervised skeleton learning. Then, in Section III, we introduce our approach, including the proposed masking strategies and model architecture for self-supervised skeleton learning. Subsequently, in Section IV, we detail our experimental setup, including the evaluation datasets, ablation studies and experimental results. Finally, we conclude and discuss future research directions in Section V.

II. Related Works

We will review related work on dynamic hand gesture recognition, focusing on skeleton-based approaches and the use of self-supervised learning methods for skeleton data.

A. Skeleton-Based Hand Gesture Recognition

Skeleton-based dynamic gesture recognition has been explored using a variety of approaches. CNNs [9], [20], [28], RNNs [11], [43] and LSTM networks [2], [30], [26]. However, GCN-based methods have become predominant for skeletal data [46], [38], [7], taking advantage of their ability to incorporate spatial connectivity between hand joints, thereby enhancing hand gesture recognition.

Temporal modeling has become increasingly significant with the work of Yan et al. [46], where hand joints are represented as a graph and spatio-temporal convolutional graph networks (STGCNs) are used to learn temporal dependencies between skeletal data from different frames in a sequence. STGCNs capture the topology of the hand through a graph adjacency matrix while learning temporal dependencies. This influential work has inspired subsequent methods to advance skeleton-based dynamic recognition of hand gestures.

Methods inspired by STGCN have explored various ways of improving performance. AS-GCN [22] introduced new modules into the ST-GCN architecture to capture actional and structural relationships, overcoming the limitations of capturing action-specific hidden joint correlations. Non-local graph convolutions [35] proposed to learn individual graphs for each sequence, allowing them to decide on connections between pairs of articulations. 2S-AAGCN [36] used a two-stream architecture to model both skeleton data and second-order information, using Adaptive GCN (AGCN) [23] to learn adjacency matrices individually for each sequence. MS-AAGCN [37] enhanced AGCN by adding a third stream for motion information. AAGCN further enhanced AGCN with a spatio-temporal attention module to focus on important articulations, images and features.

Transformers, originally designed for NLP, have been introduced as sequence models that outperform recurrent models thanks to their self-attention mechanism. Notable recent work includes STA-GCN [51], which used spatial and temporal self-attention modules to learn formable adjacency matrices, and STA-RES-TCN [17], which enhances residual temporal convolutional networks with spatio-temporal attention to focus on important frames and features. 3Mformer [44] designed a multi-order multi-mode transformer to capturing higher-order motion patterns among body joints for skeletal action recognition.

Some methods have focused on learning spatial and temporal graph adjacencies based on input graphs [22], [49], [52], [27]. By setting up an analogy between a graph and an electrical resistive network, [1] introduced physical constraints to modify joint connections by sparsing STGCNs with edge effective resistance. DG-STA [7] proposed leveraging the attention mechanism to construct dynamic temporal and spatial graphs by automatically learning node features and edges. ST-TR [33] introduced Spatial and Temporal Self-Attention modules to understand intra-frame interactions and hidden inter-frame correlations between body parts.

In this work, we leverage the strength of self-supervised learning, by learning more discriminative features for different hand poses and gestures, to improve dynamic hand gesture recognition.

B. Skeleton-based self-supervised learning

Self-supervised learning has primarily been successful in the field of image data analysis, especially due to the emergence of Masked autoencoders (MAEs) [16], demonstrating their effectiveness in a variety of applications [16], [3], [5]. Accordingly, the field of skeletal data has recently seen growing interest in exploiting the potential of self-supervised learning, particularly in the context of human gesture recognition. Researchers have explored new approaches for exploiting self-supervised learning to extract informative representations from skeletal data.

Contrastive learning methods, such as ASCAL [34] and SkeletonCLR [21], applied momentum encoders for contrastive learning using single-stream skeleton sequences. Aiming for more generalized representations, AimCLR [15] implemented an extreme data augmentation strategy to increase the number of contrastive pairs and thus improve feature extraction. To prevent overfitting and improve feature generalization for action recognition, Ms2l [24] introduced a multitasking self-supervised learning framework that focuses on the extraction of joint representations via motion

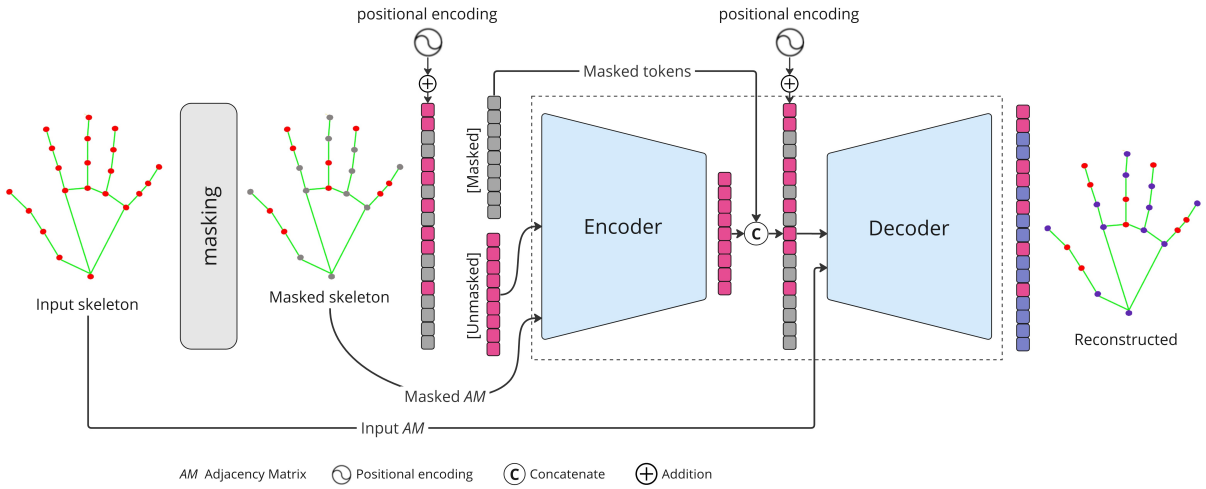

Fig. 1. **Proposed MAE for 3D hand joints reconstruction.** We mask a given ratio of hand joints, the unmasked joints are encoded by the encoder while taking into account their connectivity given by the masked adjacency matrix. The encoded joints are then concatenated with the masked tokens and passed through the decoder along with the connectivity matrix to reconstruct the masked joints.

prediction and puzzle recognition.

MAEs-based methods have been also invistigated. D-MAE [19] presented a dual-MAE approach, focusing on token completion in skeletal context, particularly relevant for robust motion capture. SkeletonMAE [45] introduced a graph-based MAE, focusing on pre-training with skeleton sequences.

Generative learning techniques such as LongT GAN [53] and P&C [40], focuses on encoder-decoder architectures for reconstructing skeleton sequences. By emphasizing the reconstruction of input sequences, they enhance feature representation by refining the encoder and decoder components.

Cloud colorization [48] explores the learning of spatial-temporal features from skeleton point clouds. This method refine feature extraction techniques by utilizing three pairs of encoder-decoder frameworks, allowing them to capture intricate spatial-temporal dependencies within the data.

However, in the specific context of hand gesture recognition, self-supervised learning remains relatively unexplored. Chen et al. [6] introduced a self-supervised framework focused on 3D hand reconstruction. SignBERT [18] proposed pre-training a hand-model-aware representation for sign language recognition.

Our work contributes to the field by advancing self-supervised skeleton learning to enhance dynamic hand gesture recognition. We propose to extend the capabilities of self-supervised skeleton learning to enhance dynamic hand gesture recognition, emphasizing the importance of leveraging prior knowledge of hand topology to refine the representations extracted from skeletal data.

III. Methodology

In this section, we first introduce the masking strategies studied in this work. Then, we present our proposed model architecture for pre-training skeleton reconstruction. Finally, we present our fine-tuning procedure for dynamic hand gesture recognition. Fig 1 illustrates the pipeline for hand skeleton reconstruction.

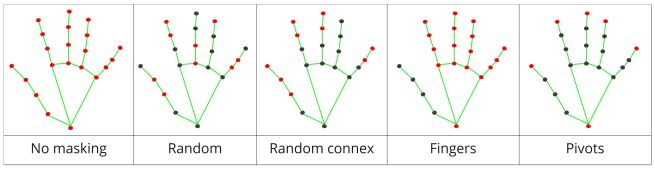

Fig. 2. Illustration of different masking strategies. Red dots: unmasked joints, gray dots: masked joints.

A. Masking Strategies

We describe various masking strategies employed in our self-supervised learning model (Fig. 2). Adapted to the hand skeleton, these strategies emphasize the importance of the different parts of the hand (fingers, pivots, tips, etc.) in learning distinctive features in various hand poses.

a) Random masking: is a widely adopted and highly effective strategy in image- and skeleton-based self-supervised learning [16], [45]. This technique involves randomly masking a number of joints in the hand skeleton.

b) Random connex masking: follows the same underlying principle as random masking, but introduces an additional constraint, namely that masked joints must be connected or adjacent in the skeleton. This constraint reinforces the structural coherence of the masked joints.

c) Fingers masking: focuses on masking the joints associated with individual fingers. The aim is for the MAE to reconstruct the masked joints based on information from the remaining fingers, enabling the model to learn the overall correlation between different fingers at different poses.

d) Pivots masking: all joints except those at the fingertips are masked. This approach is based on the assumption that, by providing the MAE with the joints located at the fingertips, it can efficiently reconstruct the other joints, regardless of the pose of the hand.

These various masking techniques play a crucial role in improving the performance of self-supervised learning

algorithms, as they enable the model to learn robust and meaningful representations from partially masked skeleton. In our experiments, we test different masking strategies and masking ratios to find the best trade-off combination.

B. Model architecture

The architecture of the MAE model is designed to process hand skeleton data. It is based on an asymmetric encoder-decoder architecture, both built upon the ViT model.

a) Encoder: Based on ViT model, we design our encoder to process skeleton data. Given the non-masked joint-level coordinates v of a hand skeleton, the encoder employs a Fourier feature mapping $\gamma(v)$ [41] to project spatial coordinates into a higher-dimensional space using sine and cosine functions of different frequencies (Equation 1). Fourier features embedding enhances the model's ability to capture spatial relationships in the skeleton data. It was specifically chosen for its unique capability to transform complex spatial dependencies into a frequency domain representation. By representing joint movements and interdependencies as frequencies, the model gains a more comprehensive understanding of the nuanced patterns in skeletal structures. This choice is motivated by the inherent ability of Fourier features to extract and encode intricate spatial dependencies in a manner that linear embedding struggles to achieve. Indeed, feeding 3D skeleton coordinates directly into the network results in a loss of fine detail, leading to less accurate representations. Conversely, pre-processing the input using Fourier feature mapping enables the network to capture detail at higher frequencies, thus improving the overall quality of the representation [41].

The Fourier feature mapping is employed to embed the 3D coordinates (x, y, z) vectors into a 256-dimensional vectors. Which are then fed into a series of ViT blocks including a self-attention mechanism and feed-forward layers to learn distinctive features in latent space for each hand pose. This architecture allows the encoder to capture complex relationships and dependencies between skeleton joints, enabling it to learn highly informative representations.

$$\begin{aligned}\gamma(v) = [&a_1 \cos(2\pi b_1^T v),\ a_1 \sin(2\pi b_1^T v),\\ &a_2 \cos(2\pi b_2^T v),\ a_2 \sin(2\pi b_2^T v),\\ &\cdots \quad , \quad \cdots \quad ,\\ &a_m \cos(2\pi b_m^T v),\ a_m \sin(2\pi b_m^T v)]^T\end{aligned} \tag{1}$$

where b are the Fourier basis frequencies used to approximate the kernel, and a are the corresponding Fourier series coefficients, resulting in a feature transformation with m distinct frequency components.

The MAE encoder is implemented on the basis of a ViT of depth 6, featuring attention mechanisms in each layer. This architecture utilizes 8 heads for multi-head attention and incorporates feed-forward networks with a dimension of 512. The embedding dimension is set to 256, encoding each 3D hand joint coordinates in a 256-element vector.

b) Decoder: The decoder is designed to reconstruct the masked joints within the skeleton data. It operates in a similar manner to the encoder but with a distinct set of parameters. The decoder first projects the encoded representations into the decoder dimension and adds positional embeddings specific to the decoder. It then concatenates the masked tokens, represented by a learnable mask token, with the encoded non-masked joints tokens. Subsequently, the decoder attends to this combined sequence using a ViT transformer. Finally, the model predicts the missing joints' coordinates.

The MAE decoder is built as a counterpart to the encoder, adopting the ViT architecture with a depth of 6. Each layer incorporates an 8-head attention mechanism for multi-head attention. The main role of the decoder is to reconstruct the joints masked in the hand skeleton data.

c) Enhancing Spatial Connectivity: Spatial connectivity between hand joints is crucial for accurate recognition of hand gestures. While ViT models intrinsically capture a certain level of spatial relationships in their attention mechanisms, the anatomical constraints of the hand skeleton can benefit from the explicit integration of adjacency matrices. In our approach, we incorporate these adjacency matrices during both the encoding and decoding phases. These matrices provide essential information that helps the model capture the connectivity between the different joints of the hand skeleton. As such, the model can make more informed predictions about missing joints, thereby improving the accuracy of the reconstruction.

This approach offers several advantages over models without adjacency information. It ensures spatial coherence and anatomical accuracy in predicted joints, crucial for gesture recognition. Additionally, it enhances the model's capability to handle complex hand configurations and poses, making it more robust for practical applications. In essence, the inclusion of adjacency matrices improves spatial modeling, enabling the attention mechanism to explicitly take into account the spatial layout of hand joints. The modified attention mechanism formula is provided in 2.

$$\text{Attention}(Q, K, V, \mathbf{A}) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} \odot \mathbf{A}\right) V \tag{2}$$

In this context, Q, K, and V represent the query, key, and value components of the original attention mechanism [42]. Additionally, A denotes the adjacency matrix, embedding spatial connectivity information between various hand joints.

Initially, we compute the raw attention scores by taking the dot product between the query and key vectors ($\frac{QK^T}{\sqrt{d_k}}$). Subsequently, we refine these scores by incorporating information from the adjacency matrix A through element-wise multiplication ($\odot$), which highlights the relevance of interconnected joints and considers hand topology. Following this, we apply the softmax function to obtain normalized attention scores. Lastly, the attention output is obtained as a weighted sum of value vectors, where the weights are determined based on the enhanced attention scores.

For the encoder, we only consider the connectivity between the non-masked joints (masked adjacency matrix),

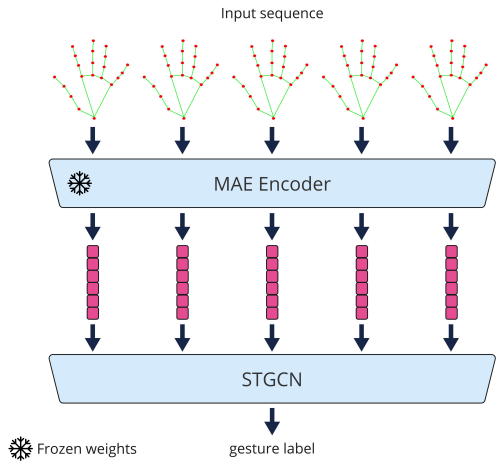

Fig. 3. In the fine-tuning phase, we exclusively use the MAE encoder with frozen weights to encode the 3D coordinates. The encoded 3D coordinates are then fed into the STGCN model to predict the corresponding gesture.

while for the decoder, the connectivity between all joints is considered (complete adjacency matrix).

d) Loss: In the pre-training phase, the Mean Squared Error (MSE) loss was employed as the main reconstruction loss function. The MSE loss is a well-established and widely used metric for evaluating the quality of reconstructions in MAE-based pre-training. It quantifies the average squared difference between the predicted and actual joints coordinates, providing a quantitative measure of the reconstruction quality. The MSE loss is defined as:

$$\mathcal{L}_{MSE} = \mathbb{E}\left[||Y - \hat{Y}||^2\right] \tag{3}$$

By minimizing the MSE loss, the MAE model attempts to make predictions as close as possible to the ground-truth coordinates. This process effectively assists the learning of distinctive representations in latent space that encompass different hand poses.

C. Fine-Tuning for Dynamic Hand Gesture Recognition

To assess the ability of the MAE model to acquire discriminative representations of the hand in various poses, we rely on the Space-Time Graph Convolutional Network (STGCN) [46] as the backbone architecture for skeleton sequence classification. The STGCN has demonstrated remarkable capabilities in learning temporal relationships, enabling it to identify complex patterns in sequential data. Additionally, an edge-attention adjacency matrix is constructed by using a learnable mask to multiply it with a predefined spatial adjacency matrix.

Given a 3D hand joint sequence, we use the MAE model's pre-trained encoder to obtain the corresponding learned representations (latent space). These learned representations serve as the basis for training the STGCN model, with performance evaluation based on recognition accuracy. The underlying process is given in Figure 3.

IV. Experiments

In the experiments section, we first present the datasets selected for our evaluation. We then conduct a series of ablation studies to assess the impact of key components of our MAE model. Finally, we present the experimental results, which include a comparative analysis with state-of-the-art methods, cross-dataset evaluations and fine-tuning with a limited labeled data.

A. Datasets

In the experiments, we assess the performance of the proposed approach using the following datasets: Briareo dataset [29], IPN Hand dataset [4], and SHREC'17 Track dataset [8]. We adhere to the respective evaluation protocols designed for each dataset to ensure consistency and fairness in our experimental evaluations.

a) Briareo [29]: released in 2019, it was primarily collected in the context of automotive applications to address driver inattention. It comprises 1440 sequences, each corresponding to 12 distinct gestures performed by 40 different subjects using their right hand. Each gesture is repeated three times. The evaluation protocol involves a division based on subjects, with 26 subjects allocated for training, 6 for validation, and 8 for testing.

b) IPN Hand [4]: includes over 4,000 gesture instances and a total of 800,000 images captured from 50 subjects. Each subject performed 21 gestures continuously, with intermittent random pauses, in a single video, with corresponding temporal segmentation provided. 13 gestures of them are designed to control pointer movements and interact with non-contact displays. In our experiments, we use the training/test split provided for evaluation.

c) SHREC'17 Track [8]: comprises 2800 sequences of gestures performed in two configurations, with a single finger and with the whole hand, each gesture being performed between 1 and 10 times by 28 participants. Depending on the number of fingers involved and the specific gesture, these sequences can be labeled into 14 or 28 classes, with accuracy evaluated accordingly. With its data on the 22-joint skeleton of the hand, this dataset offers rich information for the study of complex hand movements, making it a valuable resource for advancing gesture recognition and human-computer interaction research.

The sequences in each dataset were normalized to a uniform length. Specifically, for each dataset, we adjusted the sequence length to match the maximum number of frames per sequence, and padded shorter sequences with zeros. Consequently, sequence lengths were set at 60 for the Briareo dataset, 80 for the IPN Hand dataset and 180 for the SHREC'17 dataset.

These datasets serve as a comprehensive reference for evaluating the performance of the proposed method, and we rigorously follow the predefined evaluation protocols to guarantee reliable and consistent results.

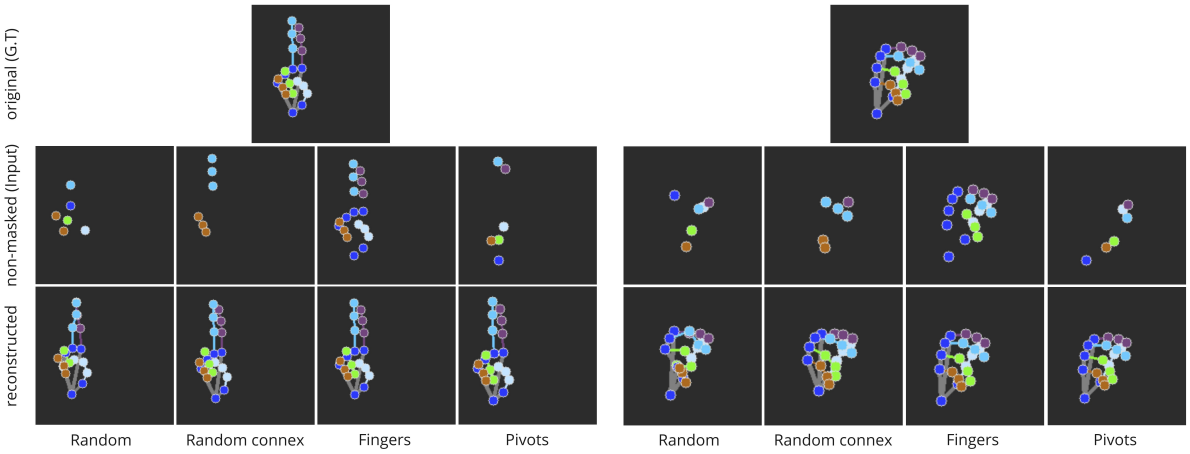

Fig. 4. Examples of reconstructed hand skeleton results when using different masking strategies.

B. Implementation Details

For datasets with no available hand landmark ground truth, namely the Briareo and IPN Hand datasets, we used the MediaPipe Hands framework [50] for hand pose estimation. This framework provides 3D (x, y, z) coordinates for 21 distinct hand joints, derived from an RGB image of the hand.

For the MAE model, key hyper-parameters include the learning rates. We selected the Adam optimizer with a learning rate of 2×10^{-4} and a weight decay of 5×10^{-2}. The learning rate is gradually reduced during training.

Regarding the STGCN, we also employed the Adam optimizer but with a learning rate of 1×10^{-3} and a weight decay of 5×10^{-2}. The learning rate is gradually reduced during training. We adopt the cross-entropy loss with label smoothing as the fine-tuning loss with a smoothing rate of 0.1 and save the best model minimizing the validation loss.

Both the pre-training and fine-tuning phases span 200 epochs. We used a batch size of 64 for pre-training and 4 for fine-tuning. Our approach was implemented using the PyTorch libraries [32].

C. Ablation Studies

To assess the effects of the main components of our MAE model, we conducted several ablation studies. Without loss of generality, we chose the Briareo dataset as our reference.

a) Feature Mapping: We began by comparing the effectiveness of Fourier feature mapping against a learned linear map implemented through a fully connected layer. The results, presented in Table I, showcase the performance difference between using linear and Fourier feature mapping in the pre-training process.

TABLE I
ABLATION STUDY ON FOURIER FEATURE MAPPING IN PRE-TRAINING.

Feature Mapping	**Linear**	**Fourier**
Accuracy	95.8%	**97.3%**

TABLE II
ABLATION STUDY ON THE INCLUSION OF THE ADJACENCY MATRIX IN PRE-TRAINING.

Adjacency Matrix	**Without**	**With**
Accuracy	94.5%	**97.3%**

TABLE III
ABLATION STUDY ON MASKING STRATEGIES.

Strategy	**Random**	**Random Connex**	**Fingers**	**Pivots**
Accuracy	**97.3%**	91.0%	90.7%	84.8%

Table I reveals that employing Fourier feature mapping significantly outperforms the learned linear mapping. The observed superiority can be attributed to the distinctive capabilities of Fourier features in capturing intricate spatial relationships among various joints in the skeleton data.

b) Inclusion of Adjacency Matrix: To assess the significance of considering spatial connectivity, we conduct an ablation study to investigate the impact of including the adjacency matrix during the pre-training of our MAE model. Table II demonstrates the influence of the adjacency matrix on recognition accuracy, highlighting its role in enhancing the MAE's ability to capture spatial correlations between different joints.

c) Masking Strategy: Fig 4 shows examples of reconstruction samples for different masking strategies. The random strategy outperforms the other strategies as shown in Table III. Conversely, pivot-based masking performs less well, as it systematically masks the same joints, preventing the model from learning the correlations and effects of masking different joints. These results demonstrate that the more random the masking, the more robust the model. This may be justified by the fact that randomness allows the model to explore all possible combinations and identify the most correlated joints. Less or non-random masking, on the other

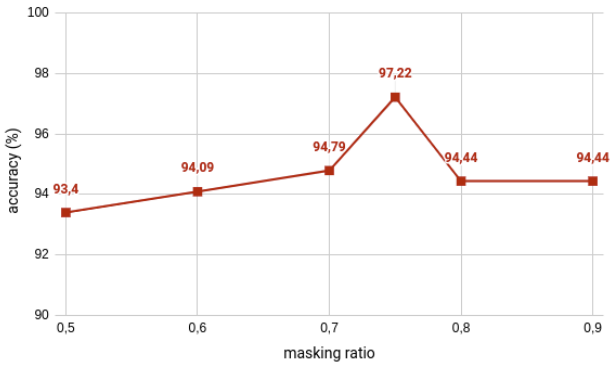

Fig. 5. Ablation study on masking ratio.

hand, systematically masks the same joints, simplifying the learning process and preventing the model from capturing the complex relationships between joints.

d) Masking Ratio: We study the effect of masking ratio while pre-training the MAE model. Figure 5 results confirm the findings on masking ratio following other work [16], [3]. Masking 75% of the joints gives the best performance. Furthermore, even when masking up to 90% of the joints, the MAE model still effectively captures and learns discriminative representations of the hand skeleton.

e) Embedding Dimension: Table IV summarizes the ablation study on the MAE decoder's embedding dimension. Utilizing an embedding dimension of 32 results in the highest accuracy (97.3%) compared to other dimensions. This finding suggests that our model can effectively learn representations with compact feature vectors.

TABLE IV
ABLATION STUDY ON EMBEDDING DIMENSION.

Embed dimension	**16**	**32**	**64**	**128**	**256**
#Parameters (M)	14	15	16	17	20
Accuracy	92.4%	**97.3%**	94.8%	96.6%	96.2%

f) Pre-training epochs: Table V shows the results of the ablation study on the pre-training scheduler. Training the model for 200 epochs yields the best results. Extending the MAE pre-training beyond 200 epochs leads to a decrease in accuracy. This drop occurs because the best model, selected based on the validation set, does not exceed the 200-epoch threshold. The key takeaway is that the optimal model performance is achieved within the first 200 epochs, and further training does not yield significant improvements, as evidenced by our ablation study results.

TABLE V
ABLATION STUDY ON THE NUMBER OF PRE-TRAINING EPOCHS.

Epochs	**50**	**100**	**150**	**200**	**300**
Accuracy	92.7%	93.4%	94.8%	**97.3%**	**97.3%**

TABLE VI
RESULTS ON BRIAREO DATASET.

Method	**Accuracy**
STGCN [46]	93.7%
STR-GCN [38]	96.6%
TBN-HGR [12]	97.2%
DG-STA [7]	90.9%
Ours	**97.3%**

TABLE VII
RESULTS ON SHREC'17 DATASET.

Method	**14 Gestures**	**28 Gestures**
STGCN [46]	92.7%	87.7%
STR-GCN [38]	93.3%	89.2%
HPEV [25]	92.5%	88.9%
DG-STA [7]	94.4%	90.7%
DD-Net [47]	94.6%	**91.9%**
STRONGER [13]	**95.0%**	-
Ours	94.1%	90.0%

D. Comparison with State-of-the-Art

In our comparative analysis, we assess the performance of our proposed approach against several state-of-the-art (SOTA) methods on different banchmark datasets including Briareo, SHREC'17 and IPN Hand datasets.

Our proposed approach consistently demonstrates competitive performance across multiple benchmark datasets. As shown in Table VII, on the SHREC'17 dataset, our method achieves a competitive accuracy score, closely matching the performance of existing SOTA methods. On the Briareo dataset, our approach surpasses the previous best result, achieving an accuracy of 97.3% as shown in Table VI. Moreover, when evaluating on the IPN Hand dataset, as presented in Table VIII, our method outperforms other SOTA techniques, achieving an accuracy of 92.8%.

The lower accuracy observed on the SHREC'17 dataset, compared to other methods, can be attributed to the mixture of coarse and fine gestures it contains. While our approach excels at capturing certain dynamic interactions, it encounters difficulties in accurately representing the nuanced spatial and temporal relationships that are crucial to accurate recognition of these diverse gestures. Further improvements to capture fine details in both spatial and temporal dimensions could enhance the model's ability to distinguish subtle variations.

TABLE VIII
RESULTS ON IPN HAND DATASET.

Method	**Accuracy**
STGCN [46]	65.1%
ResNeXt-101 [4]	86.3%
Dist-Time [14]	87.5%
Ours	**92.8%**

The results provided in Tables VI, VIII and VII indicate that our proposed approach not only achieves state-of-the-art performance but also demonstrates consistency and adaptability across various datasets. Notably, our approach demonstrates superior performance, especially considering the varying characteristics across datasets. These include the number of gestures, the number of hand joints, the sequence length, and the skeleton spatial connectivity (adjacency matrix), which are all variable across datasets.

The significant improvement in recognition accuracy achieved by using the learned representations of our MAE model as input to the STGCN model, compared to using the original 3D coordinates of the hand joints, underlines the effectiveness of our approach. This improvement in accuracy is particularly impressive because the classification model remains consistent in both cases (STGCN). The ability of our model to make hand joint features more discriminative in various hand poses demonstrates its ability to capture and represent complex patterns in the skeleton data.

E. Cross-Dataset Evaluation

In this section, we investigate the transferability of learned representations across different datasets. Specifically, we pre-train our MAE Model on dataset A and utilize the acquired weights to predict the skeleton representations of dataset B. We then fine-tune the STGCN on these transferred representations to evaluate recognition accuracy. The experimental results, presented in Table IX, demonstrate the effectiveness of cross-dataset representation learning. We chose the Briareo and IPN Hand datasets for this analysis due to their consistent skeleton annotation methods.

TABLE IX
CROSS-DATASET RECOGNITION ACCURACY.

Pre-training Dataset	Fine-tuning Dataset	Accuracy
Briareo	IPN Hand	89.7%
IPN Hand	Briareo	96.2%
Briareo + IPN Hand	Briareo	96.9%
Briareo + IPN Hand	IPN Hand	92.7%

These results underscore the robustness of our model's ability to generalize across datasets. Notably, the recognition accuracy achieved when pre-training on IPN Hand and fine-tuning on Briareo indicates that the hand poses within the IPN Hand dataset encompass a broader range, including those found in Briareo. Conversely, Briareo's hand pose dataset does not cover all the hand poses present in IPN Hand, given its automotive context viewpoint.

We also explored the impact of pre-training on a combination of datasets, specifically using both Briareo and IPN Hand, followed by fine-tuning on a single dataset. Our findings reveal that this approach does not provide a significant improvement in overall recognition accuracy over the baseline of pre-training and fine-tuning on the same single dataset (Table VI and Table VIII). Combining datasets introduces variations in terms of views and gesture dynamics, which amounts to adding noise to the data. However, the observed loss of accuracy is very marginal (0.4% for Briareo dataset and 0.1% for IPN Hand dataset), highlighting the fact that the benefits of pre-training on multiple datasets are, to a minimal extent, outweighed by dataset-specific nuances.

F. Fewer label fine-tuning

We study the ability of features learning using fewer labeled training data. We fine-tune the STGCN classification model on different randomly selected fractions of data and calculate the recognition accuracy. Table X shows the results.

TABLE X
ACCURACY WHEN FINE-TUNING ON FEWER ANNOTATED DATA.

Ratio	Briareo		IPN Hand		SHREC17 (14G)	
	seq/class	acc(%)	seq/class	acc(%)	seq/class	acc(%)
5%	4	62.2	6	64.9	7	33.6
10%	8	74.3	12	83.2	14	64.8
20%	16	87.5	24	86.0	28	55.9
50%	39	91.7	59	85.4	70	89.7
70%	55	94.8	83	91.2	98	92.0
90%	71	96.2	106	88.7	126	92.2
100%	78	97.3	118	92.8	140	94.1

The results in table X demonstrate that our MAE model can effectively learn discriminative features even when fine-tuining the classification model with limited amount of labeled data. As the ratio (%) of annotated data increases, recognition accuracy improves consistently across all datasets. Importantly, even with only 50% labeled data, the model achieves reasonable accuracy (e.g., 91.7% on Briareo), demonstrating its ability to generalize with limited supervision.

V. CONCLUSION AND FUTURE WORK

We present a new approach to self-supervised skeleton learning for dynamic hand gesture recognition. The key innovation of this research lies in the use of skeleton data and the elaboration of a tailored model architecture that efficiently captures the spatial relationships between hand joints using adjacency matrices and a ViT transformer-based encoding-decoding framework adapted to hand skeleton data.

Our experimental results showcased the state-of-the-art performance of our approach on multiple benchmark datasets, including the Briareo dataset (97.3%) and the IPN Hand dataset (92.8%). We demonstrated the model's adaptability and robustness in both single-dataset and cross-dataset settings, emphasizing its potential for real-world applications where data collection can be challenging. Additionally, we explored fine-tuning strategies with limited labeled data, revealing the model's capability to generalize even when annotated examples are scarce.

In future work, we will integrate the temporal aspect into our MAE model, allowing it to capture dynamic temporal patterns more effectively, further enhancing the discriminative power of our self-supervised learning approach for dynamic HGR. Additionally, we intend to apply our model in online settings, making it even more useful for real-world applications in human-computer interaction systems.

ACKNOWLEDGEMENT

This work is co-funded by the AI@IMT program of the Agence Nationale de la Recherche (ANR) and the region Hauts-de-France in France.

REFERENCES

[1] T. Ahmad, L. Jin, L. Lin, and G. Tang. Skeleton-based action recognition using sparse spatio-temporal gcn with edge effective resistance. *Neurocomputing*, 423:389–398, 2021.

[2] D. Avola, M. Bernardi, L. Cinque, G. L. Foresti, and C. Massaroni. Exploiting recurrent neural networks and leap motion controller for the recognition of sign language and semaphoric hand gestures. *IEEE Transactions on Multimedia*, 21(1):234–245, 2018.

[3] H. Bao, L. Dong, S. Piao, and F. Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[4] G. Benitez-Garcia, J. Olivares-Mercado, G. Sanchez-Perez, and K. Yanai. Ipn hand: A video dataset and benchmark for real-time continuous hand gesture recognition. In *25th International Conference on Pattern Recognition, ICPR 2020, Milan, Italy, Jan 10–15, 2021*, pages 4340–4347. IEEE, 2021.

[5] Y. Chen, Y. Liu, D. Jiang, X. Zhang, W. Dai, H. Xiong, and Q. Tian. Sdae: Self-distillated masked autoencoder. In *European Conference on Computer Vision*, pages 108–124. Springer, 2022.

[6] Y. Chen, Z. Tu, D. Kang, L. Bao, Y. Zhang, X. Zhe, R. Chen, and J. Yuan. Model-based 3d hand reconstruction via self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10451–10460, 2021.

[7] Y. Chen, L. Zhao, X. Peng, J. Yuan, and D. N. Metaxas. Construct dynamic graphs for hand gesture recognition via spatial-temporal attention. *arXiv preprint arXiv:1907.08871*, 2019.

[8] Q. De Smedt, H. Wannous, J.-P. Vandeborre, J. Guerry, B. L. Saux, and D. Filliat. 3d hand gesture recognition using a depth and skeletal dataset: Shrec'17 track. In *Proceedings of the Workshop on 3D Object Retrieval*, 3Dor '17, page 33–38, Goslar, DEU, 2017. Eurographics Association.

[9] G. Devineau, F. Moutarde, W. Xi, and J. Yang. Deep learning for hand gesture recognition on skeletal data. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 106–113. IEEE, 2018.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[11] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015.

[12] A. D'Eusanio, A. Simoni, S. Pini, G. Borghi, R. Vezzani, and R. Cucchiara. A transformer-based network for dynamic hand gesture recognition. In *2020 International Conference on 3D Vision (3DV)*, pages 623–632. IEEE, 2020.

[13] M. Emporio, A. Caputo, and A. Giachetti. Stronger: Simple trajectory-based online gesture recognizer. 2021.

[14] G. Fronteddu, S. Porcu, A. Floris, and L. Atzori. A dynamic hand gesture recognition dataset for human-computer interfaces. *Computer Networks*, 205:108781, 2022.

[15] 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*, volume 36, pages 762–770, 2022.

[16] 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*, pages 16000–16009, 2022.

[17] J. Hou, G. Wang, X. Chen, J.-H. Xue, R. Zhu, and H. Yang. Spatial-temporal attention res-tcn for skeleton-based dynamic hand gesture recognition. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.

[18] H. Hu, W. Zhao, W. Zhou, Y. Wang, and H. Li. Signbert: pre-training of hand-model-aware representation for sign language recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11087–11096, 2021.

[19] J. Jiang, J. Chen, and Y. Guo. A dual-masked auto-encoder for robust motion capture with spatial-temporal skeletal token completion. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5123–5131, 2022.

[20] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid. A new representation of skeleton sequences for 3d action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3288–3297, 2017.

[21] 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*, pages 4741–4750, 2021.

[22] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3595–3603, 2019.

[23] R. Li, S. Wang, F. Zhu, and J. Huang. Adaptive graph convolutional neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[24] 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*, pages 2490–2498, 2020.

[25] J. Liu, Y. Liu, Y. Wang, V. Prinet, S. Xiang, and C. Pan. Decoupled representation learning for skeleton-based gesture recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5751–5760, 2020.

[26] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot. Global context-aware attention lstm networks for 3d action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1647–1656, 2017.

[27] J. Liu, X. Wang, C. Wang, Y. Gao, and M. Liu. Temporal decoupling graph convolutional network for skeleton-based gesture recognition. *IEEE Transactions on Multimedia*, 2023.

[28] M. Liu, H. Liu, and C. Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017.

[29] F. Manganaro, S. Pini, G. Borghi, R. Vezzani, and R. Cucchiara. Hand gestures for the human-car interaction: The briareo dataset. In E. Ricci, S. Rota Bulò, C. Snoek, O. Lanz, S. Messelodi, and N. Sebe, editors, *Image Analysis and Processing – ICIAP 2019*, pages 560–571, Cham, 2019. Springer International Publishing.

[30] J. C. Nunez, R. Cabido, J. J. Pantrigo, A. S. Montemayor, and J. F. Velez. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. *Pattern Recognition*, 76:80–94, 2018.

[31] M. Oudah, A. Al-Naji, and J. Chahl. Hand gesture recognition based on computer vision: a review of techniques. *journal of Imaging*, 6(8):73, 2020.

[32] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. *In NIPS Autodiff Workshop*, 2017.

[33] C. Plizzari, M. Cannici, and M. Matteucci. Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding*, 208:103219, 2021.

[34] 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*, 569:90–109, 2021.

[35] L. Shi, Y. Zhang, J. Cheng, and H. Lu. Non-local graph convolutional networks for skeleton-based action recognition. *arXiv preprint arXiv:1805.07694*, 1(2):3, 2018.

[36] 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*, pages 12026–12035, 2019.

[37] L. Shi, Y. Zhang, J. Cheng, and H. Lu. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing*, 29:9532–9545, 2020.

[38] R. Slama, W. Rabah, and H. Wannous. Str-gcn: Dual spatial graph convolutional network and transformer graph encoder for 3d hand gesture recognition. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2023.

[39] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang. Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45(2):1474–1488, 2022.

[40] 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*, pages 9631–9640, 2020.

[41] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Ragha-

van, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[43] H. Wang and L. Wang. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 499–508, 2017.

[44] L. Wang and P. Koniusz. 3mformer: Multi-order multi-mode transformer for skeletal action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5620–5631, 2023.

[45] H. Yan, Y. Liu, Y. Wei, Z. Li, G. Li, and L. Lin. Skeletonmae: Graph-based masked autoencoder for skeleton sequence pre-training. *arXiv preprint arXiv:2307.08476*, 2023.

[46] 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*, volume 32, 2018.

[47] F. Yang, Y. Wu, S. Sakti, and S. Nakamura. Make skeleton-based action recognition model smaller, faster and better. In *Proceedings of the ACM multimedia asia*, pages 1–6. 2019.

[48] 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*, pages 13423–13433, 2021.

[49] F. Ye, S. Pu, Q. Zhong, C. Li, D. Xie, and H. Tang. Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition. In *Proceedings of the 28th ACM international conference on multimedia*, pages 55–63, 2020.

[50] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020.

[51] W. Zhang, Z. Lin, J. Cheng, C. Ma, X. Deng, and H. Wang. Sta-gcn: two-stream graph convolutional network with spatial–temporal attention for hand gesture recognition. *The Visual Computer*, 36:2433–2444, 2020.

[52] X. Zhang, C. Xu, and D. Tao. Context aware graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14333–14342, 2020.

[53] 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*, volume 32, 2018.