AG-MAE: Anatomically Guided Spatio-Temporal Masked Auto-Encoder for Online Hand Gesture Recognition

Anonymous 3DV submission

Paper ID 60

Abstract

Hand gesture recognition plays a crucial role in the domain of computer vision, as it enhances human-computer interaction by enabling intuitive, touch-free control and communication. While offline methods have made significant advances in isolated gesture recognition, real-world applications demand online and continuous processing. Skeleton-based methods, though effective, face challenges due to the intricate nature of hand joints and the diverse 3D motions they induce. This paper introduces AG-MAE, a novel approach that integrates anatomical constraints to guide the self-supervised training of a spatio-temporal masked autoencoder, enhancing the learning of 3D keypoint representations. By incorporating anatomical knowledge, AG-MAE learns more discriminative features for hand poses and movements, subsequently improving online gesture recognition. Evaluation on standard datasets demonstrates the superiority of our approach and its potential for real-world applications. Code is available at: https://github.com/lambda-xyz-01/AGMAE.

1. Introduction

Online recognition of dynamic hand gestures plays an essential role in computer vision, human-computer interaction (HCI) and virtual reality (VR) applications, enabling seamless, intuitive and natural interactions between users and machines. Unlike traditional offline gesture recognition systems [15, 20, 25], which focus on discrete segmented gestures, the framework of continuous dynamic gesture recognition requires interpreting hand movements in a continuous stream of data, enabling real-time interactions and feedbacks.

Online gesture recognition raises significant challenges due to the intricate nature of hand movements and the diverse range of motions (ROM) occurring within a continuous flow of non-segmented gestures. Unlike offline scenarios, online recognition demands precise localization of gestures within this continuous flow, necessitating accurate identification of their start and end timings. Furthermore, real-world applications necessitate real-time processing, implying rapid inference algorithms without compromising accuracy. Ensuring high precision in classification while minimizing false positives is crucial for ensuring a natural and reliable interaction experience, particularly in critical scenarios such as medical operations.

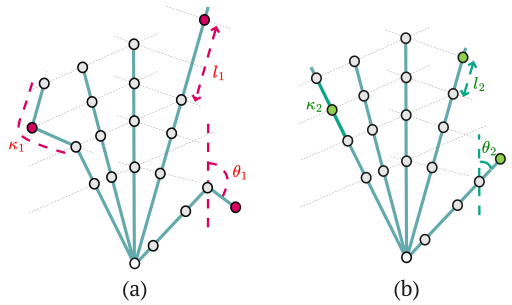


Figure 1. Hand models (a) without, (b) with anatomical constraints: joint angles (θ), bone lengths (l), finger curvature (κ).

Recent advancements in self-supervised learning have shown promise in deriving discriminative representations from unlabeled hand pose data [8, 23, 24, 42]. We argue that this approach holds significant potential, particularly in the domain of dynamic and continuous hand gesture recognition. Therefore, we propose a method that combines the power of self-supervised learning with anatomical constraints guidance to overcome the limitations inherent in traditional fully supervised approaches. By integrating self-supervised skeletal learning and anatomical information during pre-training, we aim to extract rich and discriminative representations of hand poses. As illustrated in Figure 1, anatomical constraints such as bone length, bone curvature and joint angles can be incorporated as a prior information into the learning model to ensure consistency in hand joint position estimation. Ultimately, improving the discrimination of learned 3D keypoint representations.

Existing few self-supervised learning methods [8, 23, 24] often prioritize model accuracy in matching ground

truth hand poses while neglecting anatomical correctness. By incorporating anatomical constraints into the learning process, as evidenced in prior studies on hand pose estimation [26, 38] and hand tracking [1], we demonstrate improved model capabilities in learning richer representations of various hand poses and movements.

The following are the primary contributions of our work:

- **A spatio-temporal ViT-based model with Fourier embedding:** We integrate Fourier feature embedding [40] into a spatio-temporal vision transformer (ViT) model to project spatial and temporal coordinates into a high-frequency domain. This enhancement captures intricate spatial and temporal dependencies and nuanced patterns in hand joint data, improving representation accuracy.
- **Anatomical guidance for pre-training:** We introduce anatomical constraints into the loss function to guide the pre-training of the masked autoencoder, ensuring anatomical consistency and learning discriminative features for various hand poses.

2. Related Work

2.1. Online Hand Gesture Recognition

Online gesture recognition methods extend beyond the scope of offline methods, which primarily focus on discrete, segmented gestures. In contrast, online gesture recognition involves two key tasks: segmenting the continuous data stream to identify the start and end frames of each gesture, and accurately labeling these gestures using prior information while minimizing delays and avoiding false positives. Online recognition of hand gestures has been approached through two main methods:

Full sequence-based methods: analyze an entire sequence at once to detect gesture boundaries before forwarding the identified segmented candidates to a classification module. They employ specialized heuristics based on velocity, energy, or trained networks to segment sequence and subsequently classify each frame subset. Traditional methods utilized the Histogram of Oriented Gradient (HOG) algorithm in conjunction with an SVM classifier [33]. In contrast, recent advancements have predominantly focused on time-driven models. Köpüklü et al. [28] proposed a two-model hierarchical architecture based on lightweight CNNs. Seg-LSTM [7] employs an LSTM with a specialized segmentation network, while the ST-GCN method [6] utilizes an energy-based segmentation approach with additional ad-hoc rules. The 2ST-GCN method [2, 6, 17] integrates an energy-based detection module with a fine-grained classifier for gesture/non-gesture discrimination.

Sliding window-based methods: perform continuous and simultaneous detection and labeling, often using pre-trained classifiers with fixed-size input subsequences and sliding-window models. Sliding window techniques are common, as shown by the [16] strategy, where a modified DDNet [46] is trained with segmented and resampled gestures and randomly sampled non-gestural windows. Similarly, a modified version of DeepGRU [32] demonstrated notable performance. TN-FSM [17] uses transform networks to classify 10-frame windows, while Causal TCN trains a temporal convolutional network on 20-frame windows labeled with gesture classes or non-gestures according to their intersections with the annotated ground truth [16, 17]. In addition, OO-dMVMT [12] exploits multiple temporal views of hand pose and movement to generate complete gesture descriptions.

2.2. Skeleton-based Self-Supervised Learning

Self-supervised learning has primarily been successful in image analysis, especially due to the emergence of masked autoencoders (MAEs) [22], which have been proven successful in a variety of applications [3, 9, 22]. Accordingly, the field of skeletal data has recently seen a growing interest in exploiting the potential of self-supervised learning.

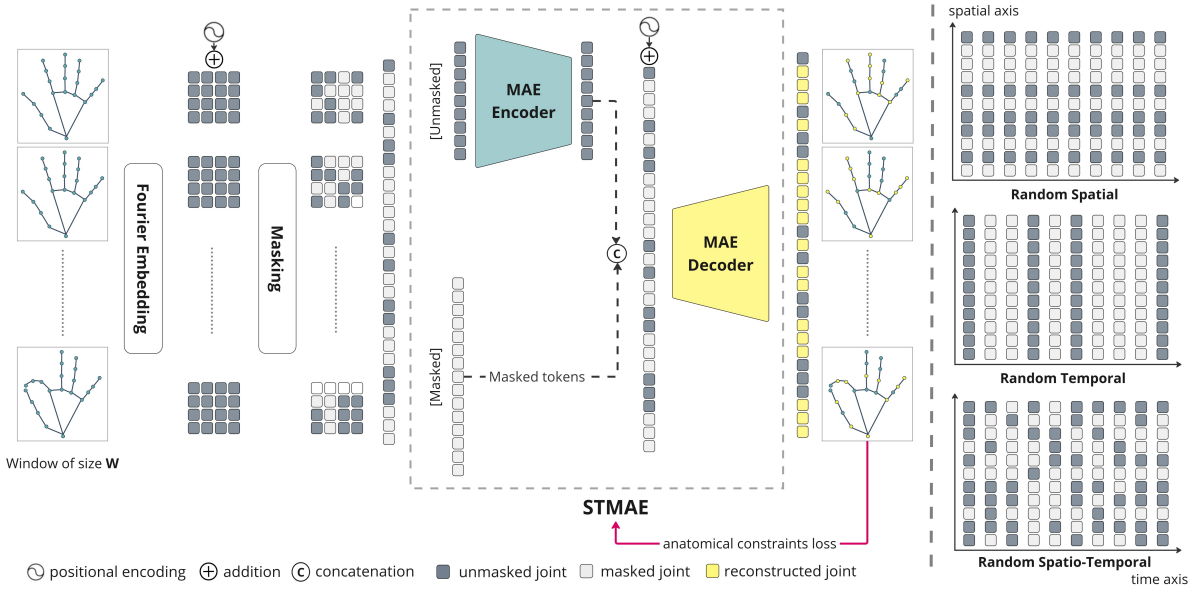
Contrastive learning methods [29, 34] apply momentum encoders for contrastive learning using single-stream skeleton sequences. Aiming for more generalized representations, AimCLR [21] implemented an extreme data augmentation strategy to increase the number of contrastive pairs and thus improve feature extraction. To prevent overfitting and improve feature generalization for action recognition, Ms2l [31] introduced a multitasking self-supervised framework that focuses on the extraction of joint representations via motion prediction and puzzle recognition.

MAE-based methods have received considerable attention. D-MAE [27] introduced a dual MAE focusing on token completion in a skeletal context, crucial for robust motion capture. Similarly, SkeletonMAE [44] proposed a graph-based MAE, emphasizing pre-training with skeleton sequences. Generative learning techniques such as LongT GAN [48] and P&C [39] emphasized encoder-decoder architectures to refine skeleton sequence representation.

Despite advances, self-supervised learning in hand gesture recognition, especially online, remains underexplored. Chen et al. [8] focused on 3D hand reconstruction, SignBERT [23] pre-trained hand-aware representations for sign language, while [24] pre-trained a MAE to encode separate individual hand poses without considering temporal correlation. Our work extends self-supervised skeleton learning to improve online gesture recognition, by incorporating spatio-temporal encoding and relying on prior knowledge and anatomical constraints to inform the learning process.

3. Methodology

We propose a comprehensive end-to-end framework for online hand gesture recognition, which is divided into two main phases. First, a spatio-temporal MAE (STMAE) is



(A) (B)

Figure 2. (A) **Proposed AG-MAE:** a ratio of joints are masked in a given window, the unmasked joints are encoded by the encoder and then concatenated with the mask tokens and passed through the decoder to reconstruct the masked joints. (B) Masking strategies.

pre-trained to encode a sequence of skeletal hand gesture frames into a robust feature representation. Subsequently, a spatio-temporal graph convolutional network (STGCN) is fine-tuned to classify gestures within a real-time data stream using the learned representations. Figure 2-A illustrates the architecture of our proposed bio-mechanically guided spatio-temporal masked autoencoder (AG-MAE).

3.1. Pretraining

In pre-training, given an input window of hand poses $X \in \mathbb{R}^{W\times N\times 3}$, where W is the number of frames, N is the number of hand joints, and 3 corresponds to the 3D coordinates (x, y, z), we first project the 3D joint coordinates into a higher-dimensional space $\mathbb{R}^d$ using a Fourier embedding map, while incorporating positional encoding to maintain spatial and temporal order. A ratio m_r (mask ratio) of joints is masked according to one of the strategies in Figure 2-B. Each joint is represented as a token of dimension d in the Fourier embedding space. The unmasked joints are processed by a ViT-based MAE encoder, mapping them to a latent space $\mathbb{R}^l$. The encoded unmasked joints are concatenated with the mask tokens and fed to the ViT-based MAE decoder to reconstruct the masked joint coordinates, producing $\tilde{X} \in \mathbb{R}^{W\times N\times 3}$. Reconstruction quality is evaluated using the mean squared error, along with the anatomical loss, which assesses the anatomical correctness of the reconstructed hand poses. This process enhances the encoding of hand poses window into a more discriminative feature space, enhancing their utility for subsequent tasks such as gesture spotting and classification.

Fourier Embedding. Fourier Feature Embedding (FFE) improves the ability of the model to capture spatial and temporal relationships between hand joints. It projects spatial and temporal coordinates into a high-frequency domain using sine and cosine functions of varying frequencies. This technique allows the model to discern nuanced patterns in 3D keypoint motions [40]. Unlike linear embeddings, which may overlook fine details, FFE preprocesses the input to capture higher-frequency details and intricate spatial dependencies, leading to more accurate representations [24, 40]. The FFE embeds the 3D coordinates $v(x, y, z)$ into a 256-dimensional space:

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

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

Positional Encoding. Positional encoding aims to preserve both spatial and temporal dimensions within the data. Specifically, a spatial positional encoding is added to each joint and maintained across all frames to retain the spatial structure. Additionally, a temporal positional encoding is applied to each frame, with the same encoding assigned to all joints within a frame to ensure temporal consistency. These encodings enable the model to effectively track and correlate spatial and temporal relationships.

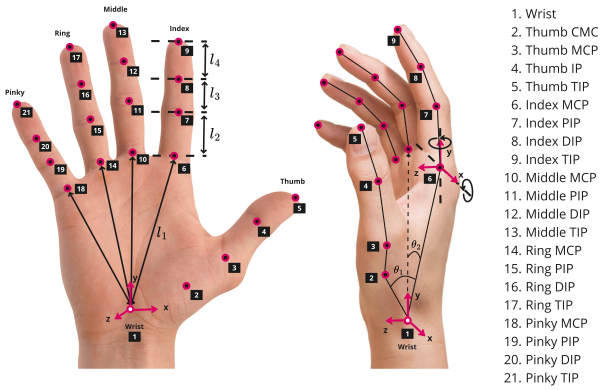


Figure 3. Hand anatomy constrains the biomechanics of hand motions, including joint angles (θ) and bone lengths (l).

Masking Strategy. Different masking strategies, illustrated in Figure 2-B, are employed to enhance the self-supervised learning model for characterizing hand poses.

- Random spatial masking: is a joint-level masking strategy that involves masking a given ratio of the same joints over time, *i.e.* the same set of random joints is masked in each frame of the sequence.
- Random temporal masking: is a frame-level masking strategy that involves masking a random number of frames in the sequences, *i.e.* all joints of the hand are masked in a given random set of frames.
- Random spatio-temporal masking: is a widely adopted and highly effective strategy in image- and skeleton-based self-supervised learning [22, 44] involving randomly masking a number of joints at both the frame- and joint-level in the sequence.

3.2. Anatomical Constraints

The hand is anatomically constrained by biomechanical limitations [11], allowing it to perform certain poses while limiting its ROM. Each joint has a specific degree of freedom (DoF) that defines its movement capabilities. For example, the index, middle, ring, and little fingers are considered planar manipulators, meaning that their DIP, PIP, and MCP joints move primarily in one plane since the DIP and PIP joints only have 1 DoF for flexion (see Figure 3). The anatomical constraints can be categorized into two primary categories: dynamic and static constraints.

Dynamic constraints can be subdivided into intrafinger and interfinger constraints. Intrafinger constraints refer to limitations on movement between different joints within the same finger [13]. For instance, Cobos et al. [11] outlined several constraints, such as the requirement that to bend the DIP joints, the PIP joints must also be bent for the index, middle, ring, and little fingers, mathematically expressed as $\theta_{DIP} = \frac{2}{3}\theta_{PIP}$. While these constraints are not rigid, individuals generally adhere to them under normal conditions, though there is some variation in the ability to control specific joints across individuals.

Interfinger constraints involve correlations between joints across different fingers, often resulting in coupled movements among fingers [30]. For example, when the pinky finger bends, the ring finger also bends to a certain extent, reflecting a proportional relationship. However, variations exist among individuals regarding these constraints. Some constraints can be overcome, while others are inherent and cannot be explicitly represented in equations [38].

Static constraints define the normal ROM for hand joints, setting limits on parameter values in models. These constraints [11], provide crucial guidelines for understanding and modeling hand biomechanics. Despite individual variations, static constraints play a significant role in defining the anatomical capabilities of the hand. The limits for each constraint can be obtained manually from measurements, from the literature (e.g., [10, 35]), or acquired in a data-driven way from 3D annotations. Two main constraints are commonly considered [11], as illustrated in Figure 3: bone lengths, reflecting intra-finger constraints, and joint angles, covering both intra- and inter-finger constraints.

For bone lengths, we define an interval $[b_i^{\min}, b_i^{\max}]$ for each bone i and penalize deviations if the length $\|b_i\|_2$, which corresponds to the Euclidean distance between the extremities of the bone at the joints, lies outside this interval. Mathematically, given a hand pose P, we define the bone length loss as:

$$\mathcal{L}_{BL}(P) = \frac{1}{N_b}\sum_{i=1}^{N_b}\mathcal{I}(\|b_i\|_2; b_i^{\min}, b_i^{\max}) \quad (2)$$

where N_b is the number of bones, and $\mathcal{I}(\|b_i\|_2; b_i^{\min}, b_i^{\max})$ is an indicator function that penalizes the bone length $\|b_i\|_2$ if it falls outside the defined interval $[b_i^{\min}, b_i^{\max}]$.

For joint angles, each joint has its specific range of freedom. For instance, as cited by Cobos et al. [11]:

$$0^\circ \le \theta_{\text{MCP}} \le 90^\circ; \quad 0^\circ \le \theta_{\text{PIP}} \le 110^\circ; \quad 0^\circ \le \theta_{\text{DIP}} \le 90^\circ.$$

We propose to compute the ranges of angles for each joint based on the data. To compute the joint angles, we first need to define a reference point relative to which the angles are measured. The wrist joint appears to be the most suitable as it has 0 DoF in the hand plane, and its movements are minimal or almost negligible.

To constrain the angles, we consider each angle independently (e.g., θ_1 between the Index MCP and Middle MCP in Figure 3) and penalize them if they lie outside the corresponding interval. This corresponds to constraining them within a box in a 2D space, where the endpoints are the min/max limits. The angles are constrained to lie within this structure by minimizing their distance to it. For angles, we consider the angles between all pairs of joints in the hand,

even those within the same hand, leading to the following definition of the loss with regards to angles:

$$\mathcal{L}_{JA}(P) = \frac{2}{N(N-1)} \sum_{i=2}^{N} \sum_{j=i+1}^{N} \mathcal{I}(\angle(\overrightarrow{J_w J_i}, \overrightarrow{J_w J_j}); a_{ij}^{\min}, a_{ij}^{\max}), \tag{3}$$

where $\angle(\overrightarrow{J_w J_i}, \overrightarrow{J_w J_j})$ is the angle between joint i and joint j considering the wrist joint (J_w) as the vertex, $a_{ij}^{min} \in A^{min}$ and $a_{ij}^{max} \in A^{max}$ are the minimum and maximum angle values between joint i and j. The number of pairs (J_w, J_i) and (J_w, J_j) where i and j are distinct joints indices is given by the binomial coefficient: $\binom{N}{2} = \frac{N(N-1)}{2}$.

Given the potential for error and inaccuracy in annotated hand gesture datasets, the lack of a common hand kinematical model across dataset and the lack of explicit equations for some dynamic angles, we propose to rely primarily on static constraints. These constraints offer a more accessible and straightforward approach to defining ROMs for individual joints and the hand as a whole. The ROM is characterized by its minimum and maximum values, determined by the angles between the various joints, with the wrist joint serving as the reference point, as well as by considerations of the distances between these joints. We argue that considering the angles between all pairs of joints in the hand provides the model with a rich feedback on the anatomical correctness of the hand.

Integrating these constraints into the loss function encourages keypoint predictions that yield valid bone lengths and valid angles, thus ensuring accurate hand anatomy. The anatomical loss can be formulated as:

$$\mathcal{L}_A(P) = \mathcal{L}_{BL}(P) + \mathcal{L}_{JA}(P) \tag{4}$$

where $\mathcal{L}_{BL}$ and $\mathcal{L}_{JA}$ denote the bone length and joint angle losses, respectively.

3.3. Model Architecture

The AG-MAE model is designed to process temporal hand skeleton data. It is based on an asymmetric encoder-decoder architecture, both built upon the ViT model [41].

Encoder. The encoder is built to encode a given window of Fourier embedded hand non-masked tokens $X^F_{\text{nmask}} = \gamma(X) \in \mathbb{R}^{N_{\text{nmask}} \times 256}$, where $X_{\text{nmask}} \in \mathbb{R}^{N_{\text{nmask}} \times 3}$ is the window of hand poses and N_{nmask} is the number of non-masked joints across the window, into a latent space $X_{\text{enc}} \in \mathbb{R}^{N_{\text{nmask}} \times d_l}$, where d_l is the latent space dimension.

The MAE encoder is implemented based on a ViT model with a depth of 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 coordinate into a 256-element vector ($d_l = 256$).

Decoder. The MAE decoder is designed to complement the encoder. It receives a complete set of tokens, which includes encoded visible patches and mask tokens (see to Figure 2). Mask tokens are shared, learned vectors that denote the presence of a missing patch that needs to be predicted. To ensure that mask tokens have location information, spatial and temporal positional embeddings are added to all tokens in this set. Subsequently, the decoder attends to this combined sets of tokens using attention mechanism and predicts the coordinates of the missing joints.

The MAE decoder is utilized exclusively during pre-training for the purpose of skeleton reconstruction, with only the encoder being employed to generate hand poses representations for downstream tasks.

Loss. During pre-training, the reconstruction loss comprises the Mean Squared Error (MSE) loss alongside the anatomical loss, represented by the ROM constraints for each joint and finger calculated given the training data. The total loss, denoted as $\mathcal{L}$, is expressed as $\mathcal{L} = \mathcal{L}_{MSE} + \lambda \mathcal{L}_A$. Here, $\mathcal{L}_{MSE} = \mathbb{E}\left[||X - \hat{X}||^2\right]$, and $\mathcal{L}_A$ signifies the anatomical loss, and λ denotes a weighting factor.

Minimizing this loss enables the MAE model to refine its predictions, striving to closely match the ground-truth coordinates while respecting anatomical correctness of hand skeleton. This iterative process facilitates the learning of discriminative representations across different hand poses within the latent space.

3.4. Fine-Tuning for Dynamic Recognition

To evaluate the effectiveness of our AG-MAE model in learning discriminative hand pose representations, we employ the spatio-temporal graph convolutional network (STGCN) [45] as the backbone architecture for classifying skeleton sequences. The STGCN excels at capturing temporal relationships, allowing it to extract complex patterns in sequential data. Additionally, it uses an edge-attention adjacency matrix constructed with a learnable mask, enhancing its ability to capture spatial dependencies.

Online Recognition. We implement a sliding-window-based model to identify the boundaries of gestures (start and end) and the gesture performed within the window of frames. The start and end are defined as the transitions between gesture classes and the non-gesture class.

The online model is based on an STGCN architecture and incorporates a classification head to predict the gesture (including the non-gesture class) within the window. Following [12], two regression heads are integrated: one for identifying the start and another for identifying the end of any detected gesture.

A cross-entropy loss is applied to the gesture class output, and an MSE loss is applied to the two regression outputs (start and end of the gesture, if any).

Offline Recognition. The offline model, on the other hand, is trained on segmented, isolated gestures. We use an STGCN model with a single classification head to output a gesture label for each segmented sequence. All sequences are padded to the dataset maximum length for uniform processing, and training is conducted using cross-entropy loss.

For both online and offline settings, given a 3D hand joint sequence, we utilize the pre-trained MAE encoder to extract the corresponding learned representations (latent space), which serve as the foundation for training the STGCN models. No masking is applied during finetuning.

4. Experimental Setup

4.1. Evaluation Protocols and Metrics

Unlike offline evaluation, which focuses mainly on recognition accuracy, online evaluation relies on more in-depth metrics to assess performance in real time, including:

Detection Rate (DR) measures the ratio of correctly detected gestures to the total number of gestures, considering temporal overlap with ground truth and duration consistency. A gesture is correctly detected if its temporal overlap exceeds 50% of the true interval, does not exceed twice the actual duration, and matches the label.

Levenshtein Accuracy (LA) captures recognition accuracy regardless of early or late detection. It's also known as minimum edit distance, meaning it measures the minimum number of single-label insertions, deletions, and substitutions needed to transform a set of labels into another.

Jaccard Index (JI) refers to the average relative overlap between ground truth and predicted labels, providing insights into the alignment of detected gestures with the ground truth gestures.

False Positive rate (FP) quantifies the ratio of false positive predictions to the total number of gestures, highlighting the ability of the model to minimize erroneous detections. Minimal false positives are desirable for robust gesture recognition systems.

Inference Time (IT) denotes the duration required for the model to perform inference and label a single frame, crucial for assessing real-time applicability.

Normalized Time to Detect (TNtD) quantifies the fraction of the sequence duration, from start to end, before the system successfully detects the gesture. Normalization aids in comparing detection performance across different sequence lengths.

4.2. Datasets

The key characteristics of the datasets used for online evaluation are given in Table 1.

Dataset	#S	#G	#J	#G/S	MeanT	StdT
SHREC21 [6]	180	17	20	3-5	77	61
IPN Hand [4]	4000	14	21	21	140	94
ODHG [14]	280	14	22	10	58	27

Table 1. Statistics for evaluation datasets: S (sequences), G (gestures), J (joints), G/S (continuous gestures per sequence), MeanT (average gesture duration), StdT (standard deviation).

SHREC'21: The SHREC'2021 Track dataset [6] meets to practical application scenarios requiring real-time gesture recognition within continuous hand movement sequences. It includes 18 gesture classes categorized as static, coarse dynamic, and fine dynamic gestures. Evaluation metrics include DR, FP rate, JI and IT.

IPN Hand: The IPN Hand dataset [4] comprises over 4,000 gesture instances from 50 subjects. Each subject executed 21 gestures continuously, interspersed with random pauses, in a single video We use the provided training/test split for evaluation. Evaluation is based on LA and IT.

ODHG: Online Dynamic Hand Gesture (ODHG) [43] is the online version of the SHREC'17 track [14], providing 280 sequences of 10 non-segmented gestures occurring sequentially and performed by 28 subjects in a continuous online environment. Evaluation is based on LA and TNtD.

Due to the variability in hand models across datasets, particularly regarding the number of joints, we propose a dataset-specific approach for inferring anatomical constraints. Specifically, we derive the ranges for these constraints, namely the minimum and maximum values for bone lengths and joint angles, based on the training set.

4.3. Implementation Details

For the STMAE model, key hyperparameters include learning rates. We employed the AdamW optimizer with a learning rate of 2×10^{-4} and weight decay of 5×10^{-2}. The learning rate is gradually reduced during training, with the biomechanical loss weighting factor (λ) set to 1.0. The window size is set to $W = 16$. Similarly, for both STGCN models, we utilized the AdamW optimizer with a learning rate of 1×10^{-3} and weight decay of 5×10^{-2}. The learning rate undergoes gradual reduction throughout training. We employed cross-entropy for training loss with label smoothing during fine-tuning, with a smoothing rate of 0.1. We use a sliding window $W = 16$ as we found it to be an optimal choice. All experiments are conducted using an `NVIDIA GeForce RTX 2080 GPU`.

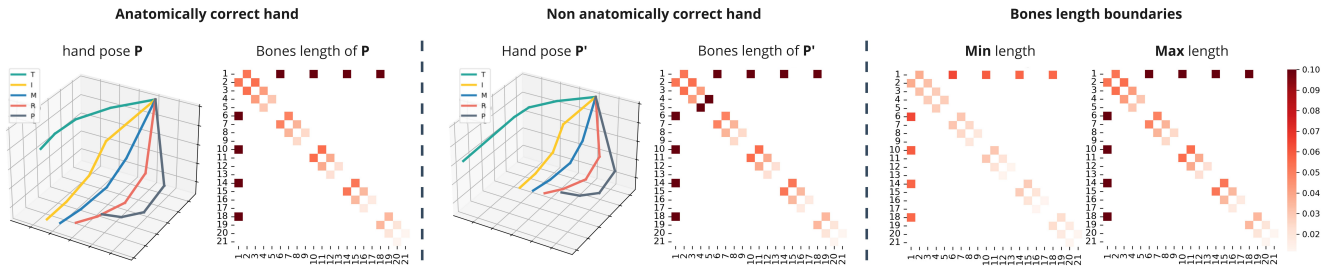


Figure 4. Left: example of an anatomically correct generated hand pose. Middle: example of a non-anatomically correct generated hand pose (thumb tip is extended beyond the normal range). Right: minimum and maximum bone lengths in the IPN Hand dataset.

5. Experimental Results

5.1. Ablation Studies

We conduct ablation studies to assess the effectiveness of various components and enhancements in our method. All experiments are conducted using the SHREC'21 dataset.

Masking Strategy. Our analysis of different masking strategies provides valuable insights (Table 2). Random spatio-temporal masking with a ratio of 0.6 ($m_r = 60\%$) proves to be the most effective, achieving 91.9% DR and a minimum FP rate of 0.033, highlighting its effectiveness for classification tasks. In contrast, random spatial masking achieves the highest JI with a ratio of 0.7, highlighting its strength in detection tasks. However, the random temporal masking shows comparatively lower performance, which can be attributed to its limited effectiveness in online gesture recognition. This reduced performance is likely due to the disruption of critical sequential patterns and temporal context that are essential for accurate real-time gesture recognition. In particular, the random nature of temporal masking can lead to masking of contiguous frames without intermediate information, disrupting the temporal flow. We argue that a guided temporal masking approach, such as one informed by joint motion, may be more effective as it reduces randomness and ensures that masking does not obscure important sequential information.

Masking strategy	Ratio	DR ↑	FP ↓	JI ↑
Random spatial	0.5	90.3%	0.0490	0.6346
	0.6	87.1%	0.0414	0.6228
	0.7	90.5%	0.0626	**0.7257**
Random temporal	0.5	74.5%	0.5331	0.5024
	0.6	80.8%	0.0516	0.5707
	0.7	81.6%	0.0753	0.5380
Random spatio-temporal	0.5	88.3%	0.0694	0.5568
	0.6	**91.9%**	**0.0330**	0.6800
	0.7	88.3%	0.0406	0.6412

Table 2. Ablation study on masking strategy and ratio in pre-training phase (SHREC'21).

Feature Embedding. We compare FFE against learned linear mapping using a fully connected layer (Table 3 - Line 1). Our experiments show that FFE significantly outperforms linear mapping. This enhancement is due to the ability of FFE to capture intricate spatial and temporal relationships among joints in skeletal data. By projecting input coordinates into a high-frequency domain, FFE allows the network to encode finer details, thereby improving the quality of learned representations.

Anatomical Loss. The inclusion of anatomical constraints significantly improves model performance, as evidenced by improvements in all evaluation metrics (Table 3 - Line 2). The anatomical loss provides critical anatomical feedback during the pre-training phase that improves model robustness without impacting inference time. This loss term helps generate anatomically correct hand poses, reducing misinterpretation of joint positions that could lead to confusion between gestures, especially in non-gesture frames that involve random hand movements.

Figure 4 illustrates the differences between correct and incorrect hand poses; the latter is shown with an exaggerated thumb extension, which is effectively penalized by the anatomical loss. In addition, the anatomical constraints adapt to different hand shapes and sizes by defining bounding ranges for bone lengths and joint angles.

Method	DR ↑	FP ↓	JI ↑	IT (ms) ↓
AG-MAE *w/o* FE	83.1%	0.082	0.573	**0.41**
AG-MAE *w/o* $\mathcal{L}_A$	84.4%	0.065	0.571	0.63
AG-MAE	**91.9%**	**0.033**	**0.680**	0.63

Table 3. Ablation studies on different components of AG-MAE model (SHREC'21).

5.2. Comparison with State-of-the-Art Methods

Offline Evaluation. We first assess our approach in an offline setting, focusing on segmented hand gesture sequences from three distinct datasets: SHREC'21, IPN Hand, and ODHG. The results, as detailed in Table 4, particularly emphasizing the critical role of self-supervised learning and

Method	Accuracy
DDNet [46]	87.8%
Stronger [16]	97.5%
AG-MAE	**98.5%**

SHREC'21 Dataset.

Method	Accuracy
ResNeXt-101 [4]	86.3%
Dist-Time [18]	87.5%
AG-MAE	**93.7%**

IPN Hand Dataset.

Method	Accuracy
G Spotter[36]	95.3%
DSTA-Net [37]	**97.0%**
AG-MAE	93.6%

ODHG Dataset.

Table 4. Offline results on evaluation datasets.

pretraining in learning spatio-temporal representations of hand skeleton data. Notably, our model achieves SOTA performance on the SHREC'21 and IPN Hand datasets.

Online Evaluation. For online evaluation, we adhere to the proposed evaluation protocol and metrics for each dataset. Table 5 shows the comparative performance of different methods on the SHREC'21 dataset. Our approach achieves SOTA results in terms of DR and FP rate. Notably, while group 4 of the original SHREC'21 paper also uses the STGCN backbone, our model, augmented with a masked autoencoder (MAE) for better representation learning, achieves an improved recognition rate of 91.9% with a notable reduction in false positives to 0.033.

However, we observe a notable decrease in the JI compared to the STGCN-based method from Group 4 (G4) of the SHREC'21 paper. This difference may stem from Group 4's use of two separate models—one for detection and one for classification—as well as their incorporation of handcrafted similarity evaluations. Such handcrafted features can be highly effective for specific gestures, contributing to their higher JI scores [6].

Method	Backbone	DR ↑	FP ↓	JI ↑	IT(ms) ↓
G1 [Shrec21] [6]	Transformer	79.2%	0.257	0.603	1.36
G2 [Shrec21] [6]	CNN	48.6%	0.927	0.277	0.41
G3 [Shrec21] [6]	GRU	75.7%	0.340	0.619	3e-6
G4 [Shrec21] [6]	STGCN	89.9%	0.066	**0.853**	0.16
Stronger [16]	CNN	90.6%	0.347	0.740	0.10
G Spotter [36]	LSTM	90.3%	0.053	0.852	-
AG-MAE	STGCN	**91.9%**	**0.033**	0.680	0.63

Table 5. Online recognition results on SHREC'21 dataset.

Method	Modality	LA ↑	IT (ms) ↓
ResNet50 [4]	RGB-Seg	33.27%	29.2
ResNet50 [4]	RGB-Flow	39.47%	43.1
ResNeXt-101 [4]	RGB-Seg	39.01%	39.9
ResNeXt-101 [4]	RGB-Flow	42.47%	53.7
TMMF [19]	RGB-Flow	68.12%	-
TSN-TSM [5]	RGB-Seg	65.27%	15.2
AG-MAE	3D keypoints	**73.93%**	19.4*

Table 6. Online evaluation on the IPN Hand dataset. (*) indicates that IT includes both keypoint extraction and inference times.

Table 6 demonstrates SOTA results of our model in terms of LA. Despite the relatively high reported inference time, it is important to note that this includes the additional time required for the extraction of 21 3D keypoints using MediaPipe [47]. Specifically, the 3D keypoints extraction contributes approximately 19.33 ms to the overall inference time, while the inference time of our model alone is on the order of $10^{-1}ms$. This suggests that the observed inference time is primarily influenced by the keypoints extraction process rather than the model itself.

For the ODHG dataset, due to the lack of a standard evaluation split protocol, we follow the authors approach and employ a random k-fold split, allocating 70% of the data for training and 30% for testing. Our model achieves an LA of 82.0% and an NTtD of 0.34. The original paper reported comparable results, with an LA of 82.2% and an NTtD of 0.21 using depth images.

Limitations. Despite the notable performance of our method, some limitations should be acknowledged. A key limitation is the trade-off between DR, FP rate, and JI depending on the masking strategy and ratio employed (see Table 2). This trade-off indicates that the optimal masking strategy and ratio may depend on the specific application requirements, highlighting the need for a balanced approach to achieve the best overall performance. Additionally, integrating the MAE with the STGCN backbone increases computational complexity, resulting in longer inference times. This may constrain the practical deployment of the model in real-time scenarios where processing speed is crucial.

6. Conclusion and Future Work

In this work, we introduce a novel framework for online hand gesture recognition combining self-supervised learning with anatomical constraints. By pre-training a spatio-temporal masked autoencoder with anatomical guidance, our approach transforms 3D hand keypoints into highly discriminative representations, enhancing performance in online gestire recognition. Comprehensive evaluations on SHREC'21, IPN Hand, and ODHG datasets shows that our method achieves SOTA results.

Future research will focus on refining adaptive masking strategies to further improve overall performance in online scenarios. Additionally, we will work on reducing model complexity to develop faster, more efficient models for deployment in resource-constrained environments.

References

[1] Andreas Aristidou. Hand tracking with physiological constraints. *The Visual Computer*, 34:213–228, 2018. 2

[2] Danilo Avola, Marco Bernardi, Luigi Cinque, Gian Luca Foresti, and Cristiano 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. 2

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2

[4] Gibran Benitez-Garcia, Jesus Olivares-Mercado, Gabriel Sanchez-Perez, and Keiji 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. 6, 8

[5] Gibran Benitez-Garcia, Lidia Prudente-Tixteco, Luis Carlos Castro-Madrid, Rocio Toscano-Medina, Jesus Olivares-Mercado, Gabriel Sanchez-Perez, and Luis Javier Garcia Villalba. Improving real-time hand gesture recognition with semantic segmentation. *Sensors*, 21(2):356, 2021. 8

[6] Ariel Caputo, Andrea Giachetti, Simone Soso, Deborah Pintani, Andrea D'Eusanio, Stefano Pini, Guido Borghi, Alessandro Simoni, Roberto Vezzani, Rita Cucchiara, et al. Shrec 2021: Skeleton-based hand gesture recognition in the wild. *Computers & Graphics*, 99:201–211, 2021. 2, 6, 8

[7] Fabio Marco Caputo, S Burato, Gianni Pavan, Théo Voillemin, Hazem Wannous, Jean-Philippe Vandeborre, Mehran Maghoumi, EM Taranta, Alaleh Razmjoo, Joseph J LaViola Jr, et al. Shrec 2019 track: online gesture recognition. In *Eurographics Workshop on 3D Object Retrieval*. The Eurographics Association, 2019. 2

[8] Yujin Chen, Zhigang Tu, Di Kang, Linchao Bao, Ying Zhang, Xuefei Zhe, Ruizhi Chen, and Junsong 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. 1, 2

[9] Yabo Chen, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Sdae: Self-distillated masked autoencoder. In *European Conference on Computer Vision*, pages 108–124. Springer, 2022. 2

[10] Fai Chen Chen, Silvia Appendino, Alessandro Battezzato, Alain Favetto, Mehdi Mousavi, and Francesco Pescarmona. Constraint study for a hand exoskeleton: human hand kinematics and dynamics. *Journal of Robotics*, 2013(1):910961, 2013. 4

[11] Salvador Cobos, Manuel Ferre, MA Sanchez Uran, Javier Ortego, and Cesar Pena. Efficient human hand kinematics for manipulation tasks. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2246–2251. IEEE, 2008. 4

[12] Federico Cunico, Federico Girella, Andrea Avogaro, Marco Emporio, Andrea Giachetti, and Marco Cristani. Oo-dmvmt: A deep multi-view multi-task classification framework for real-time 3d hand gesture classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2745–2754, 2023. 2, 5

[13] Mark R Cutkosky et al. On grasp choice, grasp models, and the design of hands for manufacturing tasks. *IEEE Transactions on robotics and automation*, 5(3):269–279, 1989. 4

[14] Q. De Smedt, H. Wannous, J.-P. Vandeborre, J. Guerry, B. Le 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*, page 33–38, Goslar, DEU, 2017. Eurographics Association. 6

[15] Naina Dhingra and Andreas Kunz. Res3atn-deep 3d residual attention network for hand gesture recognition in videos. In *2019 international conference on 3D vision (3DV)*, pages 491–501. IEEE, 2019. 1

[16] Marco Emporio, Ariel Caputo, and Andrea Giachetti. Stronger: Simple trajectory-based online gesture recognizer. 2021. 2, 8

[17] Marco Emporio, Ariel Caputo, Andrea Giachetti, Marco Cristani, Guido Borghi, Andrea D'Eusanio, Minh-Quan Le, Hai-Dang Nguyen, Minh-Triet Tran, Felix Ambellan, et al. Shrec 2022 track on online detection of heterogeneous gestures. *Computers & Graphics*, 107:241–251, 2022. 2

[18] Graziano Fronteddu, Simone Porcu, Alessandro Floris, and Luigi Atzori. A dynamic hand gesture recognition dataset for human-computer interfaces. *Computer Networks*, 205: 108781, 2022. 8

[19] Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Tmmf: Temporal multi-modal fusion for single-stage continuous gesture recognition. *IEEE Transactions on Image Processing*, 30:7689–7701, 2021. 8

[20] Mallika Garg, Debashis Ghosh, and Pyari Mohan Pradhan. Gestformer: Multiscale wavelet pooling transformer network for dynamic hand gesture recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2473–2483, 2024. 1

[21] Tianyu Guo, Hong Liu, Zhan Chen, Mengyuan Liu, Tao Wang, and Runwei Ding. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 762–770, 2022. 2

[22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2, 4

[23] Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang 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. 1, 2

[24] Omar Ikne, Benjamin Allaert, and Hazem Wannous. Skeleton-based self-supervised feature extraction for improved dynamic hand gesture recognition. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2024. 1, 2, 3

[25] Omar Ikne, Rim Slama, Hichem Saoudi, and Hazem Wannous. Spatio-temporal sparse graph convolution network

for hand gesture recognition. In *The 18th IEEE International Conference on Automatic Face and Gesture Recognition*, 2024. 1

[26] Joseph HR Isaac, Muniyandi Manivannan, and Balaraman Ravindran. Single shot corrective cnn for anatomically correct 3d hand pose estimation. *Frontiers in Artificial Intelligence*, 5:759255, 2022. 2

[27] Junkun Jiang, Jie Chen, and Yike 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. 2

[28] Okan Köpüklü, Ahmet Gunduz, Neslihan Kose, and Gerhard Rigoll. Real-time hand gesture detection and classification using convolutional neural networks. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–8. IEEE, 2019. 2

[29] Linguo Li, Minsi Wang, Bingbing Ni, Hang Wang, Jiancheng Yang, and Wenjun 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. 2

[30] John Lin, Ying Wu, and Thomas S Huang. Modeling the constraints of human hand motion. In *Proceedings workshop on human motion*, pages 121–126. IEEE, 2000. 4

[31] Lilang Lin, Sijie Song, Wenhan Yang, and Jiaying 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. 2

[32] Mehran Maghoumi and Joseph J LaViola. Deepgru: Deep gesture recognition utility. In *Advances in Visual Computing: 14th International Symposium on Visual Computing, ISVC 2019, Lake Tahoe, NV, USA, October 7–9, 2019, Proceedings, Part I 14*, pages 16–31. Springer, 2019. 2

[33] Eshed Ohn-Bar and Mohan Manubhai Trivedi. Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. *IEEE transactions on intelligent transportation systems*, 15(6):2368–2377, 2014. 2

[34] Haocong Rao, Shihao Xu, Xiping Hu, Jun Cheng, and Bin Hu. Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition. *Information Sciences*, 569:90–109, 2021. 2

[35] Chr Ryf and A Weymann. The neutral zero method—a principle of measuring joint function. *Injury*, 26:1–11, 1995. 4

[36] Junxiao Shen, John Dudley, George Mo, and Per Ola Kristensson. Gesture spotter: A rapid prototyping tool for key gesture spotting in virtual and augmented reality applications. *IEEE transactions on visualization and computer graphics*, 28(11):3618–3628, 2022. 8

[37] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In *Proceedings of the Asian conference on computer vision*, 2020. 8

[38] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *European conference on computer vision*, pages 211–228. Springer, 2020. 2, 4

[39] Kun Su, Xiulong Liu, and Eli 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. 2

[40] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020. 2, 3

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[42] Ben Veldhuijzen, Remco C Veltkamp, Omar Ikne, Benjamin Allaert, Hazem Wannous, Marco Emporio, Andrea Giachetti, Joseph J LaViola Jr, Ruiwen He, Halim Benhabiles, et al. Shrec 2024: Recognition of dynamic hand motions molding clay. *Computers & Graphics*, page 104012, 2024. 1

[43] Hazem Wannous and Jean-Philippe Vandeborre. Continuous hand gesture recognition using deep coarse and fine hand features. In *The 33rd British Machine Vision Conference–BMVC 2022*, 2022. 6

[44] Hong Yan, Yang Liu, Yushen Wei, Zhen Li, Guanbin Li, and Liang Lin. Skeletonmae: Graph-based masked autoencoder for skeleton sequence pre-training. *arXiv preprint arXiv:2307.08476*, 2023. 2, 4

[45] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 5

[46] Fan Yang, Yang Wu, Sakriani Sakti, and Satoshi Nakamura. Make skeleton-based action recognition model smaller, faster and better. In *Proceedings of the ACM multimedia asia*, pages 1–6. 2019. 2, 8

[47] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020. 8

[48] Nenggan Zheng, Jun Wen, Risheng Liu, Liangqu Long, Jianhua Dai, and Zhefeng Gong. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2