SRH Hochschule Heidelberg

Master Thesis Exposé

Skeleton-Based Hand Gesture Comparison

Author:
Amogh Shrinivas Goudar
Matriculation Number: 11037791
Wise 2023

Supervisor:
Prof. Dr. -Ing. Mehrdad Jalali
Signature:

MSc. Applied Data Science and Analytics

March 2025

SRH HOCHSCHULE HEIDELBERG
Intelligence in Learning

Table of Contents

1 Abstract

Hand gesture comparison is an essential component in various fields such as human-computer interaction, sign language interpretation, and biometric authentication. Although significant progress has been made in gesture recognition, the problem of accurately comparing hand gestures remains an open challenge. In this research, we aim to propose a transformer-based temporal modeling approach to compare 3D skeletal hand gestures by extracting embeddings and applying similarity measures. In this study, we evaluate the impact of sliding window sizes, different embedding dimensions, and visualize gesture embeddings using dimensionality reduction techniques such as t-SNE. The results of this research are expected to provide insight into optimizing gesture representations and similarity measures, contributing to advancements in gesture-based applications.

2 Introduction

Hand gesture analysis has become increasingly important in various technological and healthcare applications. In recent years, there has been a significant shift from image-based gesture recognition to 3D skeletal-based methods due to their robustness against variations in lighting conditions, camera angles, and background noise. Although gesture recognition has been extensively explored, the problem of gesture comparison remains underdeveloped. Comparing two gestures requires not only recognizing individual hand movements but also understanding their temporal dynamics and structural similarities.

Gesture comparison is essential in various domains where identifying similarities and differences between gestures is critical. For example, in sign language translation, the ability to measure the degree of similarity between gestures can improve recognition accuracy and facilitate real-time applications. Similarly, in biometric authentication, comparing hand movements enables more secure and personalized verification systems. Despite its importance, gesture comparison poses various challenges, particularly in modeling temporal dependencies and dealing with variations across different users.

2.1 Motivation

Hand gestures serve as an intuitive mode of communication in sign languages, gaming, virtual reality, and medical rehabilitation. Accurately comparing gestures is crucial for applications like gesture-based authentication systems, similarity based motion analysis, and sign language verification. The ability to compare gestures accurately is crucial for numerous real-world applications. In the domain of sign language recognition, understanding gesture similarity can significantly enhance translation accuracy. In security applications, hand gesture authentication systems can benefit from precise comparison techniques that distinguish between subtle motion differences. Moreover, in medical rehabilitation, gesture similarity analysis can help evaluate patient progress by comparing their movements with predefined reference gestures.

Despite the growing interest in gesture analysis, existing methods for gesture comparison remain limited. Current gesture comparison techniques either rely on handcrafted features or CNN/RNN-based models that fail to effectively capture long-term dependencies. Recent advancements in transformer-based architectures have demonstrated superior capabilities in learning complex temporal relationships, making them a promising choice for addressing gesture comparison challenges. This research aims to use transformer-based model architecture to improve the efficiency of gesture comparison while addressing key challenges such as intra-class variations.

2.2 Challenges

Comparing 3D skeletal hand gestures presents several fundamental challenges that must be addressed to develop a reliable and efficient approach. One of the key challenges is the inherent variability in gesture execution. The same gesture can be performed differently by different users, introducing variations in speed, articulation, and trajectory. This variability makes it difficult to establish a consistent representation for gesture comparison.

Another significant challenge lies in temporal modeling. Hand gestures are inherently dynamic, requiring an approach that can effectively capture both spatial and temporal dependencies. Many existing methods struggle with learning long-term dependencies, leading to inconsistencies in comparison accuracy. Additionally, the curse of dimensionality poses a problem, as high-dimensional skeletal data needs to be effectively transformed into a lower-dimensional yet meaningful representation for efficient comparison.

Moreover, defining a robust similarity metric for gesture comparison remains an open issue. Traditional distance-based metrics, such as Euclidean or cosine similarity, may not fully capture the intricate relationships between gestures. Instead, more sophisticated approaches that take into account contextual and sequential dependencies are needed.

2.3 Research Questions

In this section, we outline the research questions that will guide the investigation in this study:

- How can a transformer-based architecture be used to effectively capture spatial and temporal dependencies in 3D skeleton-based hand gestures?

- What is the impact of different embedding dimensions and time-window sizes on model performance?
- Which similarity metrics best capture the nuances of hand gestures?
- How can gesture embeddings be visualized to better interpret similarities and differences?

3 State of the Art

Existing techniques on gesture comparison often rely on Dynamic Time Warping (DTW) or its variants to align and compare gesture sequences. DTW can handle variations in gesture speed but it does not explicitly learn a feature representation. A study [3] highlighted that gesture comparison needs more than just recognizing individual hand movements; it also requires understanding temporal dynamics and structural similarities.

Recent research explores contrastive learning techniques to learn representations that bring similar gestures closer and push dissimilar gestures further apart. A study [1] demonstrated the effectiveness of contrastive learning in representation learning. Building on this, a study [4] has recently applied contrastive learning to skeletal-based action recognition, demonstrating improved performance in distinguishing between similar actions.

Transformers [8], initially developed for natural language processing (NLP), leverage the attention mechanism to weigh the importance of different parts of the input sequence, enabling them to capture long-range dependencies effectively. This has inspired applications in computer vision, including hand gesture analysis. A recent survey [6] provides a comprehensive overview of transformer-based approaches for video analysis, highlighting their effectiveness in capturing long-range dependencies and modeling complex temporal relationships. This research builds upon existing work in transformer-based gesture recognition [2] to improve the efficiency of gesture comparison while addressing key challenges such as intra-class variations.

Spatial Temporal Graph Convolutional Networks (ST-GCN) [9] introduces a novel approach for action recognition that leverages the spatial temporal patterns in dynamic skeletons, which are represented as graphs. STGCN applies graph-based neural networks to model dynamic skeleton sequences, allowing it to capture spatial and temporal patterns. This approach inspired our use skeleton based dataset.

FaceNet [7] introduced a deep convolutional network that learns to map face images into a compact Euclidean space, enabling efficient face verification and recognition. While primarily designed for facial analysis, its principles of embedding and similarity measurement, particularly its use of triplet loss for training, have inspired applications in other areas. This success inspired our utilization of triplet loss in the training of the gesture comparison model.

Choosing the right similarity metric is essential for gesture comparison. Common metrics include cosine similarity and Euclidean distance, but more sophisticated approaches may be needed to capture the intricate relationships between gestures. Recent work in metric learning has explored adaptive similarity metrics that learn to weight different features based on their relevance to the task. Furthermore, the study has recently proposed a learned similarity metric for action recognition that adaptively weights different features based on their relevance to the task, demonstrating improved performance compared to traditional metrics [5].

4 Expected Outcome of Thesis

The expected outcome of this thesis is a 3D skeleton-based hand gesture comparison model that distinguishes different hand gestures based on learned embeddings. The model aims to generate compact representations that enable accurate similarity measurements between gestures, optimizing the embedding space to separate distinct gesture classes. The approach is expected to improve gesture comparison performance by using transformer-based temporal modeling, and experimenting with various embedding dimensions, time-window sizes, and similarity metrics. Additionally, the research will provide insights into the visualization of gesture embeddings using t-SNE to interpret clustering behavior. The findings from this study can contribute to various applications, including sign language recognition, biometric authentication, and gesture-based human-computer interaction.

5 Methodology

5.1 Planning and Conceptualizing

The foundation of this research is built on designing a gesture comparison model that learns meaningful representations for distinguishing between different hand gestures. The model is based on a triplet learning framework, where triplets consist of an anchor, a positive, and a negative sample. The anchor and positive samples belong to the same gesture category, whereas the anchor and negative samples belong to different categories. This ensures that the model learns to minimize intra-class distances while maximizing inter-class distances in the learned embedding space.

We use the First-Person Hand Action dataset [3]. The dataset includes 3D hand pose annotations obtained using a motion capture system, which employs six magnetic sensors and inverse kinematics to infer the 3D locations of each of the 21 hand joints. It consists of 1,175 action videos covering 45 different action categories involving 26 different objects in various hand configurations. These actions are recorded in three different scenarios (kitchen, office, and social) and performed by six actors. Figure 1 shows a taxonomy of hand actions involving objects present in the dataset.

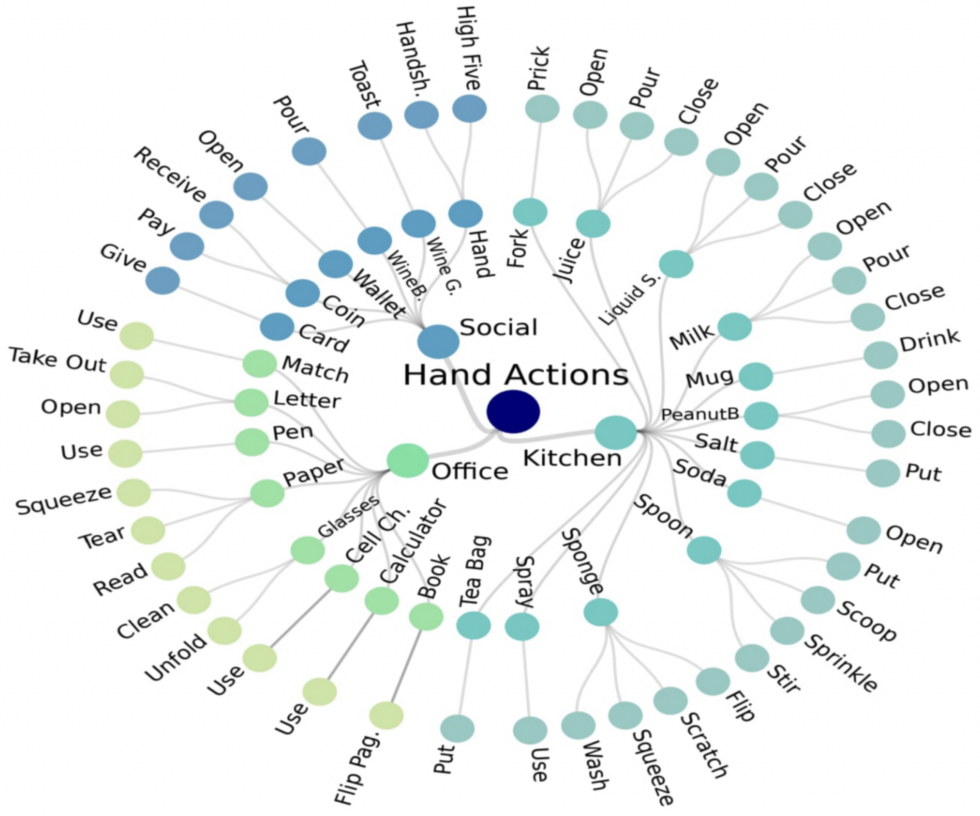


Fig. 1: First-Person Hand Action dataset

Each gesture sample is represented as a three-dimensional tensor of size T × 21 × 3, where T is the number of frames in the sequence, 21 is the number of keypoints in the hand skeleton, and 3 represents the x, y, and z coordinates. To effectively process these samples in batches, the input tensor is reshaped to b × 3 × T × 21 × 3, where b is the batch size, and 3 corresponds to the anchor, positive, and negative samples. The output of the model is structured as b × 3 × d, where d is the embedding dimension.

Figure 2 shows our approach for training gesture comparison model.

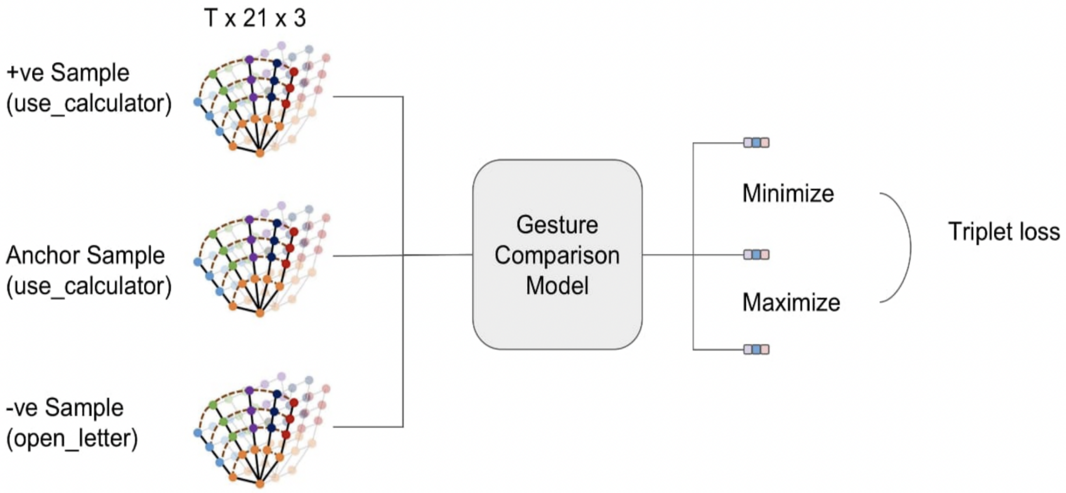


Fig. 2: Our approach for training gesture comparison model

The model utilizes a triplet loss function [7] to optimize the learned embeddings. The triplet loss ensures that the distance between the anchor and positive sample is smaller than the distance between the anchor and negative sample by a margin. This is defined as:

$$L_{\text{triplet}}(A, P, N) = \max(0, ||A - P||^2 - ||A - N||^2 + \text{margin}) \tag{1}$$

where A represents the anchor sample, P represents the positive sample, and N represents the negative sample. The margin hyperparameter controls the minimum required difference between the distances. Figure 3 illustrates the triplet loss mechanism used in this study.

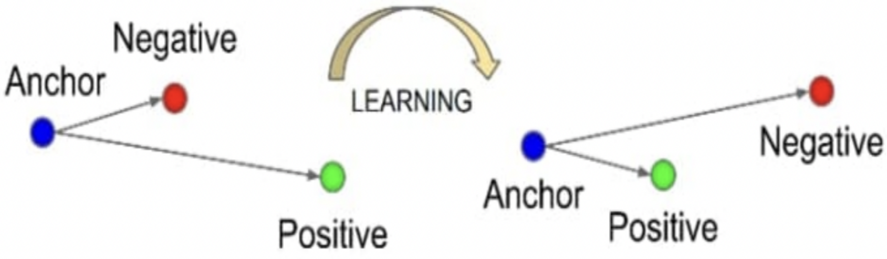


Fig. 3: Triplet Loss

We aim to develop a gesture comparison model that captures spatial and temporal dependencies in hand gestures. The model will be a transformer-based architecture designed to process sequential hand pose data while maintaining global attention across all frames. The architecture will consist of three key components: Spatial Module, Temporal Module, Embedding and Similarity Learning Module. Spatial Module module will focus on extracting meaningful spatial representations from individual frames. The joint coordinates will be mapped into a high-dimensional space using spatial embeddings, and a self-attention mechanism will be applied to integrate information across all hand joints. To model dependencies across frames, the spatially encoded representations will be flattened and concatenated into a sequence of vectors. Learnable temporal positional embeddings will be added to retain frame order information, and a temporal transformer encoder will process this sequence to effectively capture gesture dynamics over time. The embedding and similarity learning module will generate a compact gesture embedding by reducing the high-dimensional representations learned from the temporal module. Figure 4 illustrates our proposed pipeline.

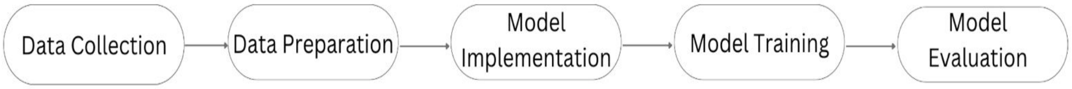


Fig. 4: Proposed Pipeline

5.2 Experimentation

In the second phase, we will design and implement an experimental pipeline using deep learning framework such as PyTorch to develop and evaluate the gesture comparison model.

The experimental phase will involve training and evaluating the model using the prepared dataset. The dataset will be divided into training and testing sets as per the dataset instructions. The model will be trained on the training set while being evaluated on the testing set to measure its performance. During training, the model's hyperparameters such as embedding dimension, distance metric, and margin for triplet loss will be experimented and tuned to find the most optimal configuration. Multiple experiments will be conducted to assess the effectiveness of different configurations, and the results will be systematically recorded for comparison. After training, the model will be evaluated on the testing set, where its performance will be assessed.

5.3 Analysis

The third phase involves analyzing the results obtained from the experimentation phase. The objective of this phase is to determine the most effective combination of model parameters and similarity metrics by comparing the outcomes of various experiments. The analysis will focus on evaluating accuracy, precision, recall, F1-score, ROC-AUC, and computational efficiency to determine which configuration produces the best performance.

Additionally, this phase will involve assessing the embedding space visualization using t-SNE or PCA to interpret clustering behavior. This will provide insight into how well the model separates similar and dissimilar gestures. The distance distribution analysis will also be examined to optimize the similarity threshold, ensuring a balanced trade-off between false positives and false negatives.

Another key part of the analysis is identifying the limitations or potential drawbacks of the proposed approach. If accuracy drops in certain scenarios, an in-depth investigation will be conducted to understand whether the cause lies in overfitting, insufficient margin tuning, or poor generalization on unseen gestures. Possible improvements will be suggested based on experimental findings to refine the approach further.

5.4 Reporting

In the final phase, the results of the study will be reported. This will involve documenting the entire research process, from planning to analysis, in a detailed report. The report will summarize the study's findings and provide recommendations for future research. It will also outline the limitations of the research and suggest ways to address them.

6 Evaluation

The evaluation of the proposed gesture comparison model is based on multiple performance metrics to ensure a comprehensive assessment. The model is optimized using a triplet-based learning approach, where the triplet loss function ensures that embeddings of similar gestures are closer together while dissimilar gestures are pushed apart.

To quantify the model's effectiveness, we measure accuracy, precision, recall, and F1-score to evaluate how well the system identifies correct gesture comparisons. Additionally, ROC-AUC analysis provides insight into the model's ability to distinguish between positive and negative gesture pairs. A visualization of the gesture embedding space using t-SNE helps analyze clustering behavior, demonstrating whether similar gestures form compact clusters while different gestures remain well-separated.

A key component of evaluation is the distance distribution analysis, where we compute the optimal threshold for gesture similarity by comparing histogram distributions of anchor-positive and anchor-negative distances. This helps minimize false positives and false negatives while maximizing separation between gesture classes.

Table 1 summarizes the evaluation criteria used to assess the model's performance.

Table 1: Evaluation metrics

Evaluation	**Distance Comparison**
True Positive (TP)	Anchor-Positive Distance $\leq$ Threshold
True Negative (TN)	Anchor-Negative Distance $>$ Threshold
False Negative (FN)	Anchor-Positive Distance $>$ Threshold
False Positive (FP)	Anchor-Negative Distance $\leq$ Threshold
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1-Score	$\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$

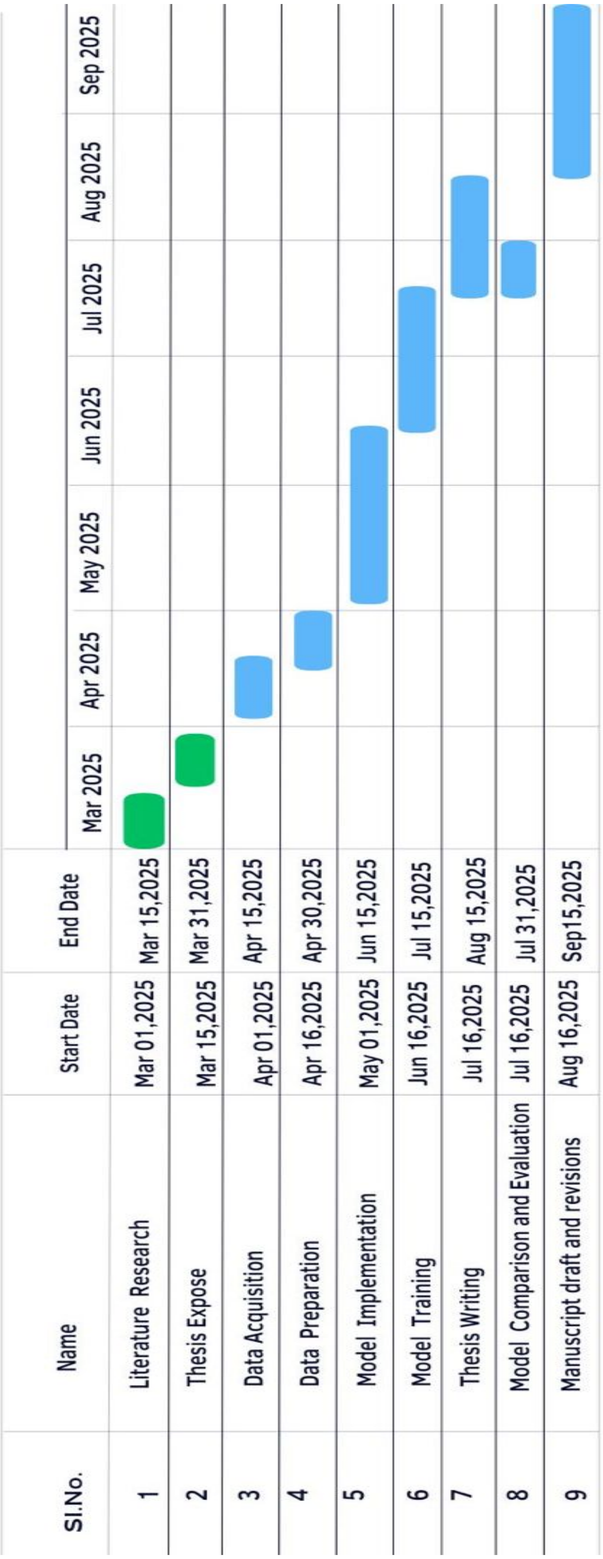
The combination of quantitative metrics and qualitative embedding space analysis ensures a robust evaluation of the proposed approach, providing a detailed understanding of its strengths and limitations.

7 Outline of Thesis

The first chapter “Introduction” will describe the motivation for this thesis and the problem statement. It also provides an overview of the research questions and challenges addressed in this study. The second chapter will be “Related Work” which will describe the state of the art. The models used for hand gesture comparison will be discussed that have been applied across various domains. The third chapter ”Methodology” describes the detailed explanation of the approach used. Focus of the fourth chapter will be “Implementation” of the model. The fifth chapter “Evaluation” corresponds to the last step. This chapter will discuss the predictions generated by the model and comparison of the model performance against the State-of-the-Art methods. In the sixth chapter “Summary and Future Work” the overall summary of the thesis, the challenges faced, and the future work will be discussed.

8 Thesis Timeline

The following timeline of the thesis work is based primarily on the corresponding chapters of the thesis. Figure 5 shows the Thesis timeline.



Sl.No.	Name	Start Date	End Date
1	Literature Research	Mar 01,2025	Mar 15,2025
2	Thesis Expose	Mar 15,2025	Mar 31,2025
3	Data Acquisition	Apr 01,2025	Apr 15,2025
4	Data Preparation	Apr 16,2025	Apr 30,2025
5	Model Implementation	May 01,2025	Jun 15,2025
6	Model Training	Jun 16,2025	Jul 15,2025
7	Thesis Writing	Jul 16,2025	Aug 15,2025
8	Model Comparison and Evaluation	Jul 16,2025	Jul 31,2025
9	Manuscript draft and revisions	Aug 16,2025	Sep15,2025

Fig. 5: Thesis Timeline

The iterative process of communicating with the supervisor, presenting the current state of the thesis will be set ahead.

References

1. Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.
2. Andrea D'Eusanio, Alessandro Simoni, Stefano Pini, Guido Borghi, Roberto Vezzani, and Rita Cucchiara. A transformer-based network for dynamic hand gesture recognition. In *2020 International Conference on 3D Vision (3DV)*, pages 623–632. IEEE, 2020.
3. Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations.
4. Xiaohu Huang, Hao Zhou, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Jingdong Wang, Xinggang Wang, Wenyu Liu, and Bin Feng. Graph contrastive learning for skeleton-based action recognition. *arXiv preprint arXiv:2301.10900*, 2023.
5. Chenglin Li, Carrie Lu Tong, Di Niu, Bei Jiang, Xiao Zuo, Lei Cheng, Jian Xiong, and Jianming Yang. Similarity embedding networks for robust human activity recognition. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(6):1–17, 2021.
6. Yang Liu, Yao Zhang, Yixin Wang, Feng Hou, Jin Yuan, Jiang Tian, Yang Zhang, Zhongchao Shi, Jianping Fan, and Zhiqiang He. A survey of visual transformers. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
7. Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.
8. 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.
9. Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition, 2018.