Performance evaluation of 3D Human Pose Estimation algorithms for skeleton-based gait recognition

Performance Evaluation of 3D Human Pose Estimation Algorithms for Skeleton-Based Gait Recognition

***Abstract*—Gait recognition is a biometric technique that identifies individuals based on unique walking patterns. Two main categories of approaches dominate this field: silhouette-based methods, which analyze a person's body shape to extract gait features, and skeleton-based methods, which focus on modeling the person's pose to represent gait movements. Silhouette-based methods are currently preferred in applications for their higher accuracy; however, skeleton-based methods are rapidly catching up, due to advancements in Human Pose Estimation (HPE) algorithms, and represent a promising upcoming alternative to silhouette-based approaches. This paper evaluates the effect of different 3D HPE algorithms on skeleton-based gait recognition, by building a gait recognition system that combines the extraction of 3D poses from RGB video data with a contrastive attention-based gait encoder for recognition. We benchmark the performance of different HPE algorithms in our system using the CASIA-B dataset, focusing on how improvements in accuracy of 3D pose data affect the Rank-1 recognition accuracy of a gait recognition algorithm. Our findings provide an analysis of suitable 3D HPE models for this task and a new benchmark result for 3D skeleton-based methods, representing a step forward in the viability of approaches based on 3D skeletons for gait recognition.**

***Index Terms*—Deep learning, gait recognition, CASIA-B, 3D human pose estimation**

I. Introduction

Gait recognition is a biometric task with the goal of recognizing individuals from the way they walk [1]. From proposed taxonomies for gait recognition [2, 3, 4], two main categories emerge among deep learning methods: silhouette-based and skeleton-based approaches.

Silhouette-based gait recognition approaches focus on segmentation of a person from a series of frames, to extract features from the segmented pixels in each frame; they generally employ Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) [5]. This category of approaches is the most used in current literature, with the first models dating back to 2008 [2].

Skeleton-based approaches focus on extracting the pose of a human subject in each frame, providing a sequence of key joint coordinates that can then be used to detect walking patterns and other gait-related features. They are also mainly based on CNNs or RNNs [6]. This category of approaches is newer, and particularly promising for its advantage of directly modeling a person's walk independently of factors that may influence the silhouette. Such factors include heavy clothing or carrying of external objects (e.g., bags or purses).

This publication is part of [BLINDED].

In the literature, silhouette-based approaches have proved to be the most effective at recognizing subjects [7], while skeleton-based ones currently lag behind in accuracy; however, the latter represent an interesting field of study because of the evolution of Human Pose Estimation (HPE) techniques in recent years [8], which contributed to their rapid rise in popularity. Furthermore, directly modeling a person's pose may prove critical for applications "in the wild", where segmenting a silhouette may be challenging due to environment complexity and image quality.

Skeleton-based approaches rely on either 2D or 3D HPE methods. The former extract the gait features starting from a sequence of 2D human poses obtained using a 2D HPE algorithm, whereas the latter extract gait features from poses obtained using 3D HPE [6, 9]. One major issue with extracting 3D poses from an RGB video source is that depth information, absent from the video frames, has to be inferred from adjacent frames [10], or from prior constraints on joints position and their relative orientation [11].

In this work, we perform quantitative comparison of 3D HPE algorithms for the purpose of performing gait recognition. To this aim, we build a gait recognition system, comprised of a pipeline that extracts a subject's poses across a video sequence using various 3D HPE algorithms interchangeably, and transforms the sequence of poses to serve as input for a state-of-the-art gait encoding approach [9]. We perform our experiments on the CASIA-B dataset [12], the most popular RGB video dataset for gait recognition tasks, and we use the Rank-1 accuracy metric under the Condition-based Matching (CME) protocol [13] to measure the effectiveness of considered algorithms.

II. Related Work

Gait Recognition. Silhouette-based methods for gait recognition have been the most popular choice for this task in the last years, due to their long-standing history and high accuracy. For instance, the Gait Energy Image (GEI) is a reliable way to extract gait information based on a sequence of silhouettes [14]. In recent years, though, skeleton-based methods have become more widespread, thanks in part to the development of highly accurate HPE algorithms. Although accuracy still cannot rival traditional approaches, skeletons are becoming increasingly viable for gait feature extraction. Yang Feng et al. [15] were among the first to extract gait features from skeletal joint coordinates using a Long Short-term Memory (LSTM)

network. Their work proved successful in recognizing subjects from the CASIA-B dataset. PoseGait [6] is one of the first methods to exploit 3D poses to perform gait recognition, with the addition of handcrafted features to improve accuracy. In the approach described in the paper, extraction of 2D poses is performed using the OpenPose framework [16]; a different algorithm is used for the 2D-to-3D lifting step, which is crucial for adding depth information to 2D poses. Haocong Rao et al. [9] developed a gait encoding approach that uses 3D skeleton data to learn gait features from 3D poses. It does so in a self-supervised way, exploiting Contrastive Attention-based Gait Encodings (CAGEs) to perform person re-identification (Re-ID). Their work opens up the possibility to perform reliable 3D pose-based gait recognition, which may serve in the future as an effective alternative to silhouette-based methods. The increasing popularity of skeleton-based approaches has also prompted the creation of datasets dedicated to these methods, such as OUMVLP-Pose [8] and Gait3D [17].

Human Pose Estimation. Human Pose Estimation (HPE) refers to the task of predicting the location of key body joints (like wrists, elbows, and hips) from visual inputs, typically from RGB images or videos. HPE methods based on deep learning can be divided into two main categories: 2D HPE and 3D HPE. In 2D HPE, the goal is to predict the 2D coordinates of keypoints in the image. Since 2014, early deep learning-based methods like DeepPose by Toshev et al. [18] introduced CNNs to predict these coordinates. Advancements introduced by later methods, such as more robust architectures, led to accurate 2D HPE systems like OpenPose [19] and HRNet [20], which became benchmarks in the field of HPE.

However, 2D pose estimation has limitations, particularly the inability to capture depth information, vital for more complex tasks like action recognition and 3D scene understanding. This limitation has driven research toward 3D HPE, which estimates the 3D coordinates of the joints adding significant value for applications such as biomechanics and animation. Direct 3D HPE methods attempt to infer 3D poses directly from RGB images, as explored by Martinez et al. [11]. Meanwhile, 2D-to-3D lifting methods first estimate 2D poses and then infer the 3D coordinates, an approach that is popular due to the availability of robust 2D pose estimators.

Monocular 3D HPE presents challenges due to the ambiguity inherent in projecting 3D poses onto 2D images. Occlusions and depth ambiguity often lead to multiple plausible 3D interpretations of a given 2D pose, as highlighted by Pavlakos et al. [21]. Despite these challenges, many methods have achieved impressive results, often using temporal information from video sequences to resolve ambiguities. Pavllo et al. [10] proposed a temporal CNN that uses sequences of 2D keypoints to predict 3D poses, leveraging the temporal continuity of human movement to improve accuracy.

III. Approach

Previous works [6, 9] have established that 3D human poses are a promising medium to extract gait features from RGB video where previously unfeasible. However, methods for 3D HPE are in constant evolution; in order to find those most suitable for this task, evaluation of different approaches is required. In this work, we build a gait recognition system by first establishing a pipeline (Figure 1) whose goal is to perform end-to-end 3D pose extraction to gait recognition from an organized set of image data. The pipeline is made up of different modules, starting with the video data loader, responsible of converting RGB video data in a format suitable for a variety of 3D HPE algorithms. Next, the 3D HPE module is responsible for extracting the 3D poses from the input data. Once the poses have been extracted, a pose transformation module manipulates the pose sequences for feeding them to the gait encoding module; the encoder proposed in [9] is used. Finally, encoding is run on the input poses, in either training, testing, or inference mode, and the Rank-1 accuracy metric is used to perform recognition.

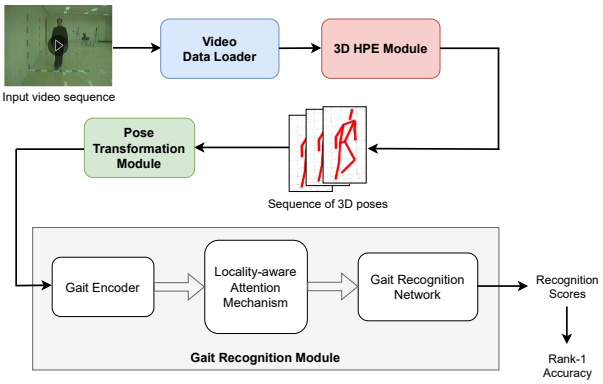

Fig. 1. The proposed pipeline for our gait recognition system, which performs recognition using 3D human poses from a single RGB video source.

A. Video Data Loader

The video data loader is responsible for loading the video data into memory in order to perform pose estimation of the subjects. At this stage, we assume only one subject is present in the input videos. The module trims the input video to the desired length for extraction of the gait features; previous works have shown that the ideal sequence length for gait extraction contains a single gait cycle, ranging from 30 to 70 frames [9], trimmed from the middle of the input video.

B. 3D Human Pose Estimation Module

The 3D HPE module is a crucial part of the pipeline. Its purpose is the extraction of 3D joint coordinates for each of the frames in the reduced frame sequence. This module enables our gait recognition system to perform an accurate identification based on the accuracy of estimated human poses, and allows for experimentation using a variety of state-of-the-art 3D HPE algoritms. The algorithms selected for this study are described in detail in Section IV-C. Previous works on gait feature extraction from 3D poses [6, 9] utilized a combination of 2D poses, obtained via the OpenPose framework [16], and an algorithm to transform these poses from 2D to 3D

coordinates. By substituting this step with pretrained end-to-end 3D HPE algorithms we aim to increase detection accuracy and model integration.

C. Pose Transformation Module

The pose transformation module is responsible for converting the poses in a format suitable for the gait encoder. To match the encoder's training data, a series of transformations must be performed on the poses, namely: conversion of the skeleton format, rotation of the poses around their root joint, translation of each pose in 3D space, and proportional scaling. The module then concatenates the poses for each frame sequence in a series of arrays for each spatial dimension (x, y, z), and identifies each sequence in a dictionary provided to the gait recognition module to compute the accuracy. Given $p \in \mathbb{R}^{j\times 3} = \{p_x, p_y, p_z\}$ as the 3D pose representation for each frame with j joints, a rotation matrix $R \in \mathbb{R}^{3\times 3}$, a translation matrix $T \in \mathbb{R}^{j\times 3} = \{T_x, T_y, T_z\}$, and a scale factor α, the behavior of the pose transformation module can be defined through the following equation:

$$p' = PT(p) = \alpha \cdot [pR(\theta_x, \theta_y, \theta_z) + T(t_x, t_y, t_z)]. \quad (1)$$

The transformed pose for each frame is a rotated, translated, and scaled version of the original pose.

The mathematical model can be simplified by only considering rotation along the z-axis and applying a fixed rotation of 90° to the x-axis. Translation parameters can also be fixed, leading to the following:

$$p' = PT(p) = \alpha \cdot [pR_x(\frac{\pi}{2})R_y(\theta)]. \quad (2)$$

The amounts of scaling and rotation, represented by α and θ, serve as hyperparameters during model training.

D. Gait Recognition Module

In order to evaluate the importance of pose data in gait recognition systems, we opted to utilize the gait encoder, the locality-aware attention mechanism, and the recognition network provided by the CAGEs approach [9], which represents the leading 3D pose-based gait encoder in the literature. We integrate such method in our gait representation module. The gait encoder works in a self-supervised way, learning to reconstruct skeletons in reverse joint order as well as other pretext tasks. The gait recognition network uses Rank-1 recognition accuracy in the CME evaluation protocol (described in [13]) to recognize the subjects. The gait encoder from the CAGEs approach can be trained and utilized on datasets that do not provide a pose ground truth, such as CASIA-B, as well as RGB data captured in the wild.

IV. Experiments

A. Dataset

For our experiments, we used the CASIA-B gait recognition dataset [12]. It includes 124 walking subjects, captured by 11 simultaneous views rotated by 18° increments, with the subjects walking from 0° to 180° relative to the camera. The subjects were captured in three different modalities: walking with normal clothes (nm), while carrying a bag (bg), and wearing heavy clothing (cl). The dataset offers multiple low-resolution video sequences for each modality and subject. This allows for one or more videos to be used as a gallery, and the rest to be used as probes. Training protocols for gait recognition on the CASIA-B dataset usually select the first 74 subjects for training and the remaining 50 for testing, although different training splits can and have been considered [5]. The CASIA-B dataset was chosen because of the high number of subjects and sequences, as well as considering that it is one of very few gait recognition datasets that provides RGB video data using different modalities, required for our video-to-gait encoding pipeline. The need for image data prevented us from using other popular datasets for gait recognition, such as OUMVLP-Pose [8] and Gait3D [17], which only provide pose or mesh data.

B. Evaluation Protocol

To evaluate the performance of our gait recognition system on the CASIA-B dataset, we adopt the protocol used by the authors of the gait encoder [9], known as Condition-based Matching Evaluation (CME).

For each testing subject, a set of gait sequences from one of the modalities (nm, bg, cl) is used as the probe, while the gallery consists of one or more sequences from the nm modality. Rank-1 accuracy, the most widely used metric in gait recognition, indicates the system's ability to identify the correct individual by comparing each probe sequence to all the gallery sequences and matching the probe with the highest-scoring identity using a similarity metric (e.g., Euclidean distance or cosine similarity).

The CME protocol, on the other hand, defines evaluation as being conducted under both single-condition (i.e., the gallery and probe sets are from the same condition with no appearance changes) and cross-condition settings (i.e., the probe set is from the normal condition while the gallery is from the bag or clothes condition), focusing on cross-condition generalization. The CME protocol evaluates the Rank-1 accuracy across different viewing angles and modalities, assessing the system's robustness to variations. We report the average accuracy across all the angles for each gallery-probe pair.

C. Model Selection and Training

For the 3D HPE module, various state-of-the-art models were considered. The main requirement was the ability to perform monocular 3D HPE on short video sequences. Therefore, models that utilize the temporal component for HPE were preferred, since this approach has proved effective on video sources compared to single-frame methods [10].

One such model that was selected for evaluation is GAST-Net [22], which exploits a Graph Attention Spatio-temporal CNN to extract 3D poses. GAST-Net is one of the best-performing algorithms on the 3D HPE dataset Human3.6M [23], with an average Mean Per-Joint Position Error (MPJPE)

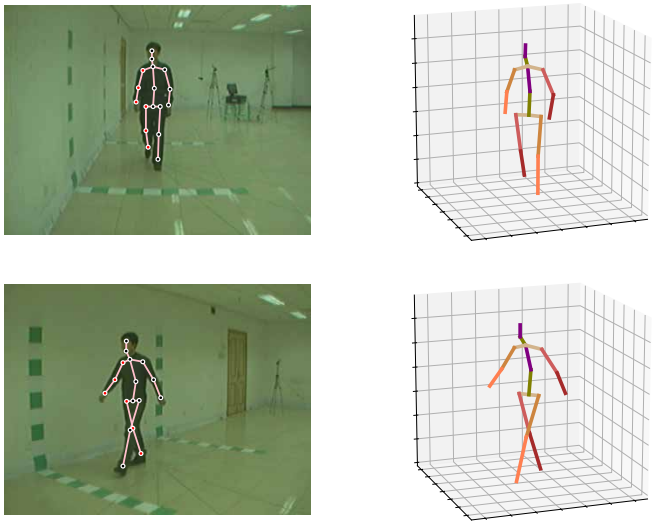

Fig. 2. Samples of 3D skeleton poses extracted from the CASIA-B training set using GAST-Net [22].

of 44.9 mm. Pose estimation samples on CASIA-B using this algorithm can be seen in Figure 2.

Three more models were then selected, namely PoseFormer [24], PoseFormerV2 [25], and MHFormer [26]. They are all based on the transformer architecture [27], which has proved effective for HPE among many other computer vision tasks.

The models were implemented inside the 3D HPE module of our pipeline, and initial experimental runs were performed. The gait encoding module was trained in each run based on the poses output from the considered HPE model. In the end, the MHFormer model was selected for parameter optimization due to its lower complexity (for implementation) and computational requirements.

D. Results

The results for the evaluation of our gait recognition system under the CME setup using the different selected HPE models are reported in Table I. The score for CAGEs is taken from [9], which was the highest CME score at the time, although it produced very similar results in our tests. The scores for the selected models, except for MHFormer, are taken from the best run considering non-exhaustive sets of hyperparameters. An in-depth hyperparameter search was conducted for MHFormer, for which we report the results obtained with the best set of hyperparameters: $\alpha = 500, \theta = 144$, with medium-sample (MT) training split.

From the analysis, it can be noted that MHFormer consistently outperforms the other models across most scenarios, achieving the highest accuracy in the nm-nm, bg-bg, cl-cl, and bg-nm categories with scores of 61.3, 44.2, 39.9, and 41.4, respectively. However, PoseFormer stands out in the cl-nm category, where it achieves the best score of 27.1. PoseFormerV2 critically underperforms in cross-modality scenarios, proving unsuitable for this task. GAST-Net shows weaker but consistent performance compared to CAGEs.

TABLE I

COMPARISON BETWEEN THE GAIT ENCODING APPROACH IN [9] AND THE MODELS USED IN THIS STUDY. THE REPORTED RANK-1 ACCURACY SCORES FOR THE SELECTED MODELS ARE EVALUATED USING THE CONDITION-BASED MATCHING (CME) PROTOCOL.

Model	nm-nm	bg-bg	cl-cl	cl-nm	bg-nm
CAGEs [9]	54.3	37.5	31.9	27.0	36.3
GAST-Net [22] + CAGEs	44.1	28.4	25.8	24.1	30.1
PoseFormer [24] + CAGEs	48.7	30.4	24.1	**27.1**	33.4
PoseFormerV2 [25] + CAGEs	55.4	33.3	37.1	1.04	1.6
MHFormer [26] + CAGEs	**61.3**	**44.2**	**39.9**	23.5	**41.4**

Overall, being CAGEs the most accurate gait encoding approach based on 3D poses in the state of the art under the CME protocol [9], our improvements to the recognition pipeline using MHFormer represent a new benchmark result for gait recognition based on 3D human poses.

Although state-of-the-art silhouette-based approaches provide very high Rank-1 accuracy (e.g., the authors of GaitSet [5] report a nm-nm accuracy of 96%), skeleton-based methods are worth further research due to their promising way of modeling a person's walking patterns. Our results demonstrate that the quality of 3D skeleton data can have positive impacts on gait recognition accuracy, paving the way for 3D pose-based gait recognition approaches that may rival silhouette-based ones in the future.

V. CONCLUSION

In this paper, we presented a comprehensive evaluation of different 3D HPE algorithms for their application to gait recognition tasks. Our proposed gait recognition system is comprised of a pipeline that can integrate multiple 3D HPE algorithms to extract gait features from monocular RGB video sequences and subsequently feed them into a state-of-the-art gait encoding model from 3D poses, with appropriate geometrical transformations.

Although accuracy scores do not yet rival silhouette-based approaches, experimental results on the CASIA-B dataset demonstrated highest-ever Rank-1 accuracy for 3D skeleton-based methods in multiple modalities under the CME protocol. Our results prove that improvements in gait recognition based on 3D poses are achievable thanks to advancements in HPE algorithms, and they may soon catch up to more traditional gait encoding approaches.

Future work will focus on further refining the pipeline by integrating more advanced pose transformation methods and collecting data for new, diverse datasets captured in uncontrolled environments to test more in depth the capabilities of our approach under real-world conditions. We plan on conducting a case study of our pipeline on advanced automotive biometric systems, by enriching training and testing data with RGB video data captured from real-world automotive applications.

REFERENCES

[1] A. Kale, N. Cuntoor, B. Yegnanarayana, A.N. Rajagopalan, and R. Chellappa. "Gait analysis for human identification". In: *Proceedings of the 3rd Conference*

in Audio- and Video-Based Biometric Person Authentication. 2003, pp. 706–714.

[2] Alireza Sepas-Moghaddam and Ali Etemad. "Deep gait recognition: A survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1 (2023), pp. 264–284.

[3] Jasvinder Pal Singh, Sanjeev Jain, Sakshi Arora, and Uday Pratap Singh. "Vision-based gait recognition: A survey". In: *IEEE Access* 6 (2018), pp. 70497–70527.

[4] Chuanfu Shen, Shiqi Yu, Jilong Wang, George Q. Huang, and Liang Wang. *A comprehensive survey on deep gait recognition: Algorithms, datasets and challenges*. 2023. arXiv: 2206.13732.

[5] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. *Gaitset: regarding gait as a set for cross-view gait recognition*. 2018. arXiv: 1811.06186.

[6] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. "A model-based gait recognition method with body pose and human prior knowledge". In: *Pattern Recognition* 98 (2020), p. 107069.

[7] Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. "Gait lateral network: learning discriminative and compact representations for gait recognition". In: *Proceedings of the 16th European Conference in Computer Vision*. Springer-Verlag, 2020, pp. 382–398.

[8] Weizhi An et al. "Performance evaluation of model-based gait on multi-view very large population database with pose sequences". In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 2.4 (2020), pp. 421–430.

[9] Haocong Rao et al. "A self-supervised gait encoding approach with locality-awareness for 3d skeleton based person re-identification". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.10 (2022), pp. 6649–6666.

[10] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. *3d human pose estimation in video with temporal convolutions and semi-supervised training*. 2019. arXiv: 1811.11742.

[11] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. *A simple yet effective baseline for 3d human pose estimation*. 2017. arXiv: 1705.03098.

[12] Shiqi Yu, Daoliang Tan, and Tieniu Tan. "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition". In: *Proceedings of the 18th International Conference on Pattern Recognition*. IEEE, 2006.

[13] Zheng Liu, Zhaoxiang Zhang, Qiang Wu, and Yunhong Wang. "Enhancing person re-identification by integrating gait biometric". In: *Neurocomputing* 168 (2015), pp. 1144–1156. ISSN: 0925-2312.

[14] J. Han and Bir Bhanu. "Individual recognition using gait energy image". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.2 (2006), pp. 316–322.

[15] Yang Feng, Yuncheng Li, and Jiebo Luo. "Learning effective gait features using Lstm". In: *Proceedings of the 23rd International Conference on Pattern Recognition*. IEEE, 2016.

[16] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. "Openpose: realtime multi-person 2d pose estimation using part affinity fields". In: *CoRR* abs/1812.08008 (2018). arXiv: 1812.08008.

[17] Jinkai Zheng, Xinchen Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei. *Gait recognition in the wild with dense 3d representations and a benchmark*. 2022. arXiv: 2204.02569.

[18] Alexander Toshev and Christian Szegedy. "Deeppose: human pose estimation via deep neural networks". In: *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2014.

[19] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. *Realtime multi-person 2d pose estimation using part affinity fields*. 2017. arXiv: 1611.08050.

[20] Jingdong Wang et al. *Deep high-resolution representation learning for visual recognition*. 2020. arXiv: 1908.07919.

[21] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. *Ordinal depth supervision for 3d human pose estimation*. 2018. arXiv: 1805.04095.

[22] Junfa Liu, Juan Rojas, Zhijun Liang, Yihui Li, and Yisheng Guan. *A graph attention spatio-temporal convolutional network for 3d human pose estimation in video*. 2020. arXiv: 2003.14179.

[23] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. "Human3.6m: large scale datasets and predictive methods for 3d human sensing in natural environments". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.7 (2014), pp. 1325–1339.

[24] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. *3d human pose estimation with spatial and temporal transformers*. 2021. arXiv: 2103.10455.

[25] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. *Poseformerv2: exploring frequency domain for efficient and robust 3d human pose estimation*. 2023. arXiv: 2303.17472.

[26] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. *Mhformer: multi-hypothesis transformer for 3d human pose estimation*. 2022. arXiv: 2111.12707.

[27] Ashish Vaswani et al. *Attention is all you need*. 2023. arXiv: 1706.03762.