PoseFormerV2: Exploring Frequency Domain for Efficient and Robust 3D Human Pose Estimation

Qitao Zhao[1]* Ce Zheng[2] Mengyuan Liu[3] Pichao Wang[4] Chen Chen[2]

[1]Shandong University [2]Center for Research in Computer Vision, University of Central Florida [4]Amazon Prime Video
[3]Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School

qitaozhao@mail.sdu.edu.cn cezheng@knights.ucf.edu nkliuyifang@gmail.com
pichaowang@gmail.com chen.chen@crcv.ucf.edu

Abstract

*Recently, transformer-based methods have gained significant success in sequential 2D-to-3D lifting human pose estimation. As a pioneering work, PoseFormer captures spatial relations of human joints in each video frame and human dynamics across frames with cascaded transformer layers and has achieved impressive performance. However, in real scenarios, the performance of PoseFormer and its follow-ups is limited by two factors: (a) The length of the input joint sequence; (b) The quality of 2D joint detection. Existing methods typically apply self-attention to **all frames** of the input sequence, causing a huge computational burden when the frame number is increased to obtain advanced estimation accuracy, and they are not robust to noise naturally brought by the limited capability of 2D joint detectors. In this paper, we propose PoseFormerV2, which exploits a compact representation of lengthy skeleton sequences in the frequency domain to efficiently scale up the receptive field and boost robustness to noisy 2D joint detection. With minimum modifications to PoseFormer, the proposed method effectively fuses features both in the time domain and frequency domain, enjoying a better speed-accuracy trade-off than its precursor. Extensive experiments on two benchmark datasets (i.e., Human3.6M and MPI-INF-3DHP) demonstrate that the proposed approach significantly outperforms the original PoseFormer and other transformer-based variants. Code is released at* `https://github.com/QitaoZhao/PoseFormerV2`.

1. Introduction

3D human pose estimation (HPE) aims at localizing human joints in 3-dimensional space based on monocular videos (without intermediate 2D representations) [23,25] or 2D human joint sequences (referred to as 2D-to-3D lifting approaches) [5,17,33,39]. With the large availability of 2D human pose detectors [6,24] plus the lightweight nature of 2D skeleton representation of humans, lifting-based methods are now dominant in 3D human pose estimation. Compared to raw monocular videos, 2D coordinates of human joints in each video frame are much more memory-friendly, making it possible for lifting-based methods to utilize a long joint sequence to boost pose estimation accuracy.

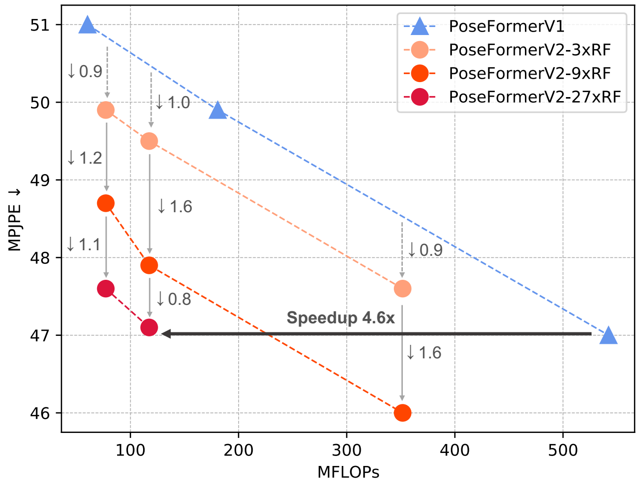

Figure 1. Comparisons of PoseFormerV2 and PoseFormerV1 [41] on Human3.6M [12]. RF denotes Receptive Field and $k\times$RF indicates that the ratio between the full sequence length and the number of frames as input into the spatial encoder of PoseFormerV2 is k, *i.e.*, the RF of the spatial encoder is expanded by $k\times$ with a few low-frequency coefficients of the full sequence. The proposed method outperforms PoseFormerV1 by a large margin in terms of speed-accuracy trade-off, and the larger k brings more significant improvements, *e.g.*, 4.6× speedup with the k of 27.

Transformers [32] first gain huge success in the field of natural language processing (NLP) [3, 7] and then extend their capacity to the computer vision community, becoming the *de facto* approach for several vision tasks, *e.g.*, image classification [8, 18, 31], object detection [4, 42] and video recognition [1,2,38]. The discreteness of human joint representation and the requirement for long-range temporal

[*]Work was done while Qitao was an intern mentored by Chen Chen.

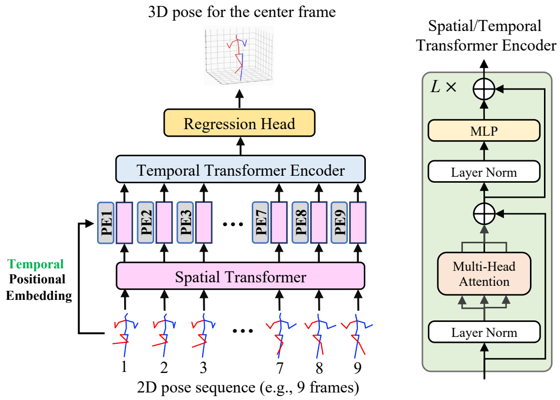

Figure 2. Overview of PoseFormerV1. PoseFormerV1 mainly consists of two modules: the spatial transformer encoder and the temporal transformer encoder. The temporal encoder of PoseFormerV1 applies self-attention to all frames given a 2D joint sequence for human motion modeling.

Table 1. The computational cost and **performance drop** brought by replacing ground-truth 2D detection with CPN [6] 2D pose detection for the SOTA transformer-based methods. The performance drop is reported on Human3.6M dataset (Protocol 1) [12]. RF: Receptive Field, sharing the same meaning as that in Fig. 1.

Method		Seq. Length	GFLOPs	Perform. **Drop** (mm)
PoseFormerV1 [41]	ICCV'21	81	1.36	13.0
StridedTransformer [14]	TMM'22	243	1.37	15.2
MixSTE [40]	CVPR'22	81	92.46	16.5
MHFormer [15]	CVPR'22	81	3.12	11.8
P-STMO [29]	ECCV'22	243	1.74	13.5
PoseFormerV2 (9×RF)		81	0.35	**8.2**
PoseFormerV2 (27×RF)		81	**0.12**	9.7

dependency modeling in a skeleton sequence make transformers an excellent fit for lifting-based human pose estimation. Previous works [14, 15, 29, 40, 41] have adopted transformers as the backbone for 3D human pose estimation and shown promising results.

As the pioneering work among transformer-based methods, PoseFormer [41] factorizes joint sequence feature extraction into two stages (see Fig. 2) and outperforms traditional convolution-based approaches. First, all joints within each frame are linearly projected into high-dimensional vectors (*i.e.*, joint tokens) as input into the spatial transformer encoder. The spatial encoder builds up inter-joint dependencies in single frames with the self-attention mechanism. In the second stage, joint tokens of each frame are combined as one frame token, serving as input to the temporal encoder for human motion modeling across all frames in sequence. More details are included in Sec. 3.1.

Despite its capacity, the performance of PoseFormer (and other transformer-based methods) is limited by two crucial factors. **(a)** The length (number of frames) of the input 2D skeleton sequence. State-of-the-art transformer-based methods typically use extremely long sequences to obtain advanced performance, *e.g.*, 81 frames for PoseFormer [41], 243 frames for P-STMO [14] and 351 frames for MHFormer [15]. However, densely applying self-attention to such long sequences is highly computationally expensive, *e.g.*, the single-epoch wall-time training cost of 3-frame PoseFormer is ∼5 minutes while for 81-frame PoseFormer the cost surges to ∼1.5 hour on an RTX 3090 GPU. **(b)** The quality of 2D joint detection. 2D joint detectors inevitably introduce noise due to bias in their training dataset and the temporal inconsistency brought by the single-frame estimation paradigm. For example, PoseFormer achieves 31.3mm MPJPE (Mean Per Joint Position Error) using the **ground-truth** 2D detection on the Human3.6M dataset [12]. This result drops significantly to 44.3mm when the clean input is replaced by the CPN [6] 2D pose detection. In practice, the long-sequence inference may be unaffordable for hardware deployment on resource-limited devices such as AR/VR headsets and high-quality 2D detection is hard to obtain. More quantitative results about the efficiency to process long sequences and the robustness to noisy 2D joint detection of existing transformer-based methods are available in Table 1.

Driven by these practical concerns, we raise two important research questions:

- *Q1: How to efficiently utilize long joint sequences for better estimation precision?*
- *Q2: How to improve the robustness of the model against unreliable 2D pose detection?*

Few works have tried to answer either of these two questions by incorporating hand-crafted modules, *e.g.*, the downsampling-and-uplifting module [9] that only processes a proportion of video frames for improved efficiency, the multi-hypothesis module [15] to model the depth ambiguity of body parts and the uncertainty of 2D detectors. *However, none of them manages to find a single solution to these two questions simultaneously, and even worse, a paradox seemingly exists between solutions to the questions above,* e.g.*, multiple hypotheses* [15] *improve robustness but bring additional computation cost (see also Table 1).*

In this paper, we present our initial attempt to *"kill" two birds with one stone*. With restrained modifications to the prior art PoseFormer, we show that the appropriate form of representation for input sequences might be the key to answering these questions simultaneously. Specifically, we shed light on the barely explored frequency domain in 3D HPE literature and propose to encode the input skeleton sequences into low-frequency coefficients. The insight behind this representation is surprisingly simple: On the one hand, low-frequency components are enough to represent the entire visual identity [34, 37] (*e.g.*, 2D images in image compression and joint trajectories in this case), thus removing the need for expensive all-frame self-attention; On the other, the low-frequency representation of the skeleton sequence itself filters out high-frequency noise (jitters and outliers) [19, 20] contained in detected joint trajectories.

We inherit the spatial-temporal architecture from PoseFormer but force the spatial transformer encoder to only "see" a few central frames in a long sequence. Then we complement "short-sighted" frame-level features (the output of the spatial encoder) with global features from low-frequency components of the complete sequence. Without resorting to the expensive frame-to-frame self-attention for all time steps, the temporal transformer encoder is reformulated as a Time-Frequency Feature Fusion module.

Extensive experiments on two 3D human pose estimation benchmarks (*i.e*., Human3.6M [12] and MPI-INF-3DHP [21]) demonstrate that the proposed approach, dubbed as **PoseFormerV2**, significantly outperforms its precursor (see Fig. 1) and other transformer-based variants in terms of speed-accuracy trade-off and robustness to noise in 2D joint detection. Our **contributions** are three-fold:

- To the best of our knowledge, we are the first to utilize a frequency-domain representation of input joint sequences for 2D-to-3D lifting HPE. We find this representation an ideal fit to concurrently solve two important issues in the field (*i.e*., the efficiency to process long sequences and the robustness to unreliable joint detection), and experimental evidence shows that this approach can easily generalize to other models.
- We design an effective Time-Frequency Feature Fusion module to narrow the gap between features in the time domain and frequency domain, enabling us to strike a flexible balance between speed and accuracy.
- Our PoseFormerV2 outperforms other transformer-based methods in terms of the speed-accuracy trade-off and robustness on Human3.6M and achieves the state-of-the-art on MPI-INF-3DHP.

2. Related Work

Our method is built on conceptually simple PoseFormer [41], and we aim at improving its efficiency to operate long sequences and its robustness to noisy joint detection from a frequency-domain perspective. Therefore, here we mainly focus on this line of works (transformer-based methods) in 2D-to-3D lifting HPE and introduce applications of frequency domain representations in computer vision literature, especially in skeleton-based tasks that are most related to lifting-based 3D HPE.

2.1. Transformer-based 3D Human Pose Estimation

PoseFormer [41] is the first work to adopt the vision transformer as the backbone network in lifting-based 3D human pose estimation, and it outperforms previous CNN-based methods by a large margin. Though being competitive, Zhang *et al*. [40] point out that the spatial-then-temporal paradigm of PoseFormer may neglect distinct temporal patterns for each joint, and propose to adopt alternate spatial-temporal transformer layers for fine-grained joint-specific feature extraction. MHFormer [15] further incorporates task-related prior knowledge into transformers for 3D HPE. Specifically, 2D-to-3D lifting is an inverse problem where more than one reasonable solutions exist, therefore they generate multiple hypotheses to model ambiguous body parts and uncertainty in joint detectors, achieving advanced performance. Inspired by the progress of Masked Image Modeling (MIM) in image classification [11, 35, 36], P-STMO [29] applies Masked Joint Modeling to 3D HPE with self-supervised learning.

Another line of works [9, 14] improves the efficiency of transformer-based methods. Taking advantage of the temporal redundancy in 2D joint sequences, StridedTransformer [14] replaces the parameter-heavy fully-connected layers with strided convolutions. Einfalt *et al*. [9] claim that the per-frame 2D joint detection is even more computationally expensive than lifting models themselves and propose to downsample input video frames with a fixed interval and adopt the 2D joint detector and lifting model only on these sampled frames. While being more efficient than previous works, aforementioned methods [9, 14] reduce participants in self-attention along the temporal dimension utilizing only the consistency in adjacent video frames rather than from a global view, and therefore they may suffer from a considerable performance drop.

2.2. Frequency Representation in Vision

Since the human visual system is more sensitive to low-frequency components of images, traditional image compression algorithms, *e.g*., JPEG [27] and JPEG 2000 [30], reduce memory cost to store 2D images by allocating more storage budget to low-frequency Discrete Cosine Transform (DCT) coefficients of the image. Following the same logic, [37] proposes to adaptively remove uninformative channels of DCT components for 2D images to boost image classification efficiency. More recently, some works [10, 28] propose to replace the costly self-attention mechanism with frequency transforms that can be accelerated by their fast algorithms. GFN [28] proposes to efficiently mix visual tokens with learnable frequency filters, and AFNO [10] further improves the performance of token mixer in the frequency domain with operator learning. Moreover, Wang *et al*. [34] utilize low-frequency Fast Fourier Transform (FFT) components to compress vision transformers.

Skeleton-based tasks are more relevant to our work that takes 2D skeleton sequences as input. In the human motion prediction literature, previous works [19, 20] transform the skeleton sequence from the time domain into DCT coefficients to encode human dynamics as compared to static joint coordinates. They observe that discarding a few high-frequency coefficients does not necessarily bring a performance drop but even improves the smoothness of predicted future motions. However, frequency-domain representations of 2D joint sequences have not yet been explored in lifting-based 3D human pose estimation.

Our approach is inspired by these former attempts of applying frequency transforms to vision tasks but from a different view. We include more details about our motivations to choose the DCT coefficient representation in Sec. 3.2.1.

3. Method

PoseFormer [41] facilitates 3D human pose estimation by factoring sequence feature extraction into two stages, *i.e.*, the spatial encoder and temporal encoder, which has proved to be effective. However, it suffers from a huge computational burden when the length of input sequences is increased, and is sensitive to noisy joint detection. In this section, we introduce the details of PoseFormerV2 which utilizes a frequency representation of the input sequence to overcome two aforementioned problems.

3.1. Preliminaries of PoseFormerV1 [41]

We start by giving a brief overview of PoseFormerV1 (see Fig. 2), laying the basis for the discussions on its improvements. PoseFormerV1 consists of two main modules, the spatial encoder for single-frame joint correlation modeling and the temporal encoder for cross-frame human motion modeling. Given an input 2D skeleton sequence $\mathbf{x} \in \mathbb{R}^{F \times J \times 2}$, where F denotes the sequence length and J denotes the joint number of the human representation. First, coordinates of all joints of a person in each frame are linearly projected to a c-dimensional vector (*i.e.*, joint embedding) denoted as $\mathbf{z}_0 \in \mathbb{R}^{F \times J \times c}$. A learnable spatial positional embedding $\mathbf{E}_{SPos} \in \mathbb{R}^{1 \times J \times c}$ [8] is added to $\mathbf{z}_0$ to encode joint-dependent information.

Spatial transformer encoder builds up spatial dependencies for joint embeddings of each frame $\mathbf{z}_0^i \in \mathbb{R}^{1 \times J \times c}$ individually with the self-attention mechanism. In this stage, the number of tokens fed into each transformer block is J. The output of the spatial transformer encoder of L layers for the i-th frame is denoted by $\mathbf{z}_L^i \in \mathbb{R}^{1 \times J \times c}$. Then per-frame representations are flattened and concatenated as input $\mathbf{Z}_0 \in \mathbb{R}^{F \times (J \cdot c)}$ into the temporal transformer encoder.

Temporal transformer encoder. Similarly, the input $\mathbf{Z}_0$ is added with a learnable temporal positional embedding $\mathbf{E}_{TPos} \in \mathbb{R}^{F \times (J \cdot c)}$ to encode index-dependent information for each frame. The temporal encoder with M transformer layers densely models frame-to-frame dependencies across the whole sequence, and its output is denoted by $\mathbf{Z}_M \in \mathbb{R}^{F \times (J \cdot c)}$. In this stage, the token number for each transformer layer is F, which is the input sequence length.

Regression head. To estimate the 3D pose of the central frame in sequence, a simple 1D convolution is used to gather temporal information and a linear projection outputs the final pose representation $\mathbf{y} \in \mathbb{R}^{1 \times (J \cdot 3)}$.

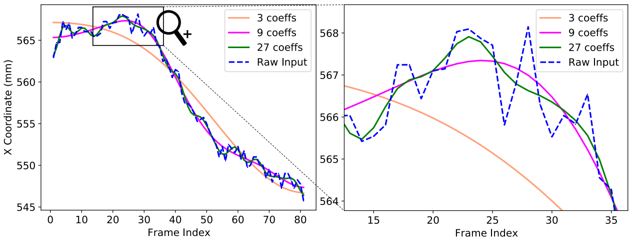

Figure 3. A randomly selected example of the CPN-detected [6] joint trajectory in Human3.6M [12] and its reconstructions with first 3, 9, and 27 DCT coefficients (81 in total). Note that even with only the first 3 coefficients, the reconstructed (orange) curve captures the overall characteristics of the raw input, and is smoother.

Limitations of PoseFormerV1. Modeling joint dependencies within each frame and human motions across frames with transformer layers is straightforward. While such dense modeling brings advanced estimation accuracy, it is computationally unfriendly due to the quadratic computation growth of self-attention with respect to the token number (*i.e.*, the joint number in the spatial encoder and the sequence length in the temporal encoder) especially when the input sequence length is increased. Although the token number for spatial transformer layers (*i.e.*, the joint number) is independent of the frame number, it is worth noting that the sequence length implicitly affects the computational budgets of the spatial encoder in real scenarios because of the limited parallelization ability of GPUs. In addition to the efficiency issue, PoseFormerV1 is sensitive to the quality of input 2D joint detection (experimental evidence is available in Table 1 and Sec. 4.3).

In the following, we present an alternative solution to overcome the limitations of PoseFormerV1 with the frequency-domain representation of the input sequence.

3.2. PoseFormerV2

3.2.1 Frequency Representation of Skeleton Sequence

Motivation. We propose to transform the input skeleton sequence into the frequency domain with Discrete Cosine Transform (DCT) and utilize only a portion of low-frequency coefficients. DCT coefficients encode multiple levels of temporal information for the input time series. Specifically, low-frequency coefficients encode its rough contour while high-frequency ones encode its details, *e.g.*, jitters or sharp changes. To better illustrate our motivation to choose this representation, we provide an 81-frame example of the CPN-detected [6] joint trajectory of action “Directions” in the Human3.6M [12] dataset and its reconstructions with first 3, 9, and 27 DCT coefficients respectively (see Fig. 3). As the number of kept DCT coefficients increases, the reconstructed trajectory becomes closer to the raw input but less smooth. Note that with only 3 DCT coefficients (denoted by the orange curve), the overall trend of the original trajectory is captured, and with 9 and 27

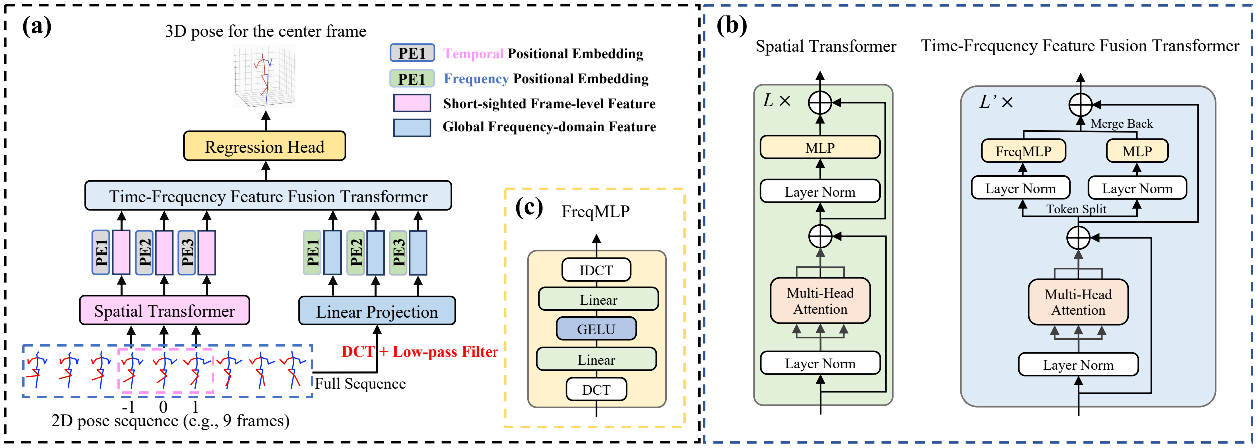

Figure 4. (a) Overview of **PoseFormerV2**. (b) Spatial Transformer and Time-Frequency Feature Fusion Transformer. (c) FreqMLP (Frequency Multi-Layer Perceptron). To exemplify, in (a) we use 3 central frames (index -1, 0, and 1) for fine-grained frame-level feature extraction and the first 3 DCT coefficients of the full 9-frame sequence for global frequency-domain feature extraction. Therefore, the effective number of frames as input to the spatial encoder and temporal encoder is reduced compared to PoseFormerV1 (3,6 *vs*. 9,9).

coefficients (pink and green curves), the characteristics of the raw sequence are better preserved while high-frequency noise (zig-zags) is removed. These observations motivate us to exploit a few highly informative low-frequency DCT components of the input joint sequence as the *compact and denoised sequence representation* in our work. With such representation, we significantly reduce the effective length of the sequence as input and promote the robustness of our model against the noise contained in 2D joint detection. We include a formal introduction to DCT in supplementary.

3.2.2 Architecture

In this part, we introduce the architecture of the proposed approach, PoseFormerV2 (see Fig. 4 for an overview).

Spatial transformer encoder. Given a 2D skeleton sequence $\mathbf{x} \in \mathbb{R}^{F \times J \times 2}$ (preferably a long sequence, *e.g.*, F is 81), we first sample F' (typically $F' \ll F$) frames around the sequence center (the frame of index 0 in Fig. 4 (a)), denoted by $\mathbf{x}' \in \mathbb{R}^{F' \times J \times 2}$, as input to the spatial encoder. The output of the spatial encoder is denoted by $\mathbf{z}^{Time} \in \mathbb{R}^{F' \times (J \cdot c)}$ (frame-level features in the time domain). The design of the spatial encoder directly follows PoseFormerV1.

Low-frequency DCT coefficients. $\mathbf{z}^{Time}$ is referred to as "short-sighted" because its receptive field (F') is restricted in comparison to the entire sequence length (F). To efficiently exploit the long-range human dynamics of the original sequence, we resort to its frequency-domain representation. We first convert the full sequence $\mathbf{x} \in \mathbb{R}^{F \times J \times 2}$ into DCT coefficients, denoted by $\mathbf{C} \in \mathbb{R}^{F \times J \times 2}$. Then we keep only the first N ($\ll F$) coefficients $\mathbf{C}' \in \mathbb{R}^{N \times J \times 2}$ using a low-pass filter for every joint trajectory, where temporal information of the original sequence is largely maintained and high-frequency noise is removed. Low-frequency coefficients $\mathbf{C}'$ are flattened and linearly projected to $\mathbf{z}^{Freq} \in \mathbb{R}^{N \times (J \cdot c)}$ (the embedding of frequency coefficients). $\mathbf{z}^{Freq}$ is summed with a learnable frequency positional embedding $\mathbf{E}_{FPos}$ (like $\mathbf{E}_{TPos}$ in PoseFormerV1). Features from both the time domain and frequency are concatenated together, formulated as

$$\mathbf{z} = [\mathbf{z}^{Time}; \mathbf{z}^{Freq}], \tag{1}$$

fed to the Time-Frequency Feature Fusion module.

Time-Frequency Feature Fusion. We adopt transformer layers for cross-frame temporal dependency modeling as in PoseFormerV1. Compared to PoseFormerV1 which entirely extracts features in the time domain, the proposed method fuses features from both the time domain and frequency domain. To narrow the gap between the two domains, we introduce simple modifications to vanilla transformer layers. **(1)** Time-domain and frequency-domain features share self-attention but use separate feed-forward networks; **(2)** We apply FreqMLP (Frequency Multi-Layer Perceptron) in the feed-forward networks for time-domain features $\mathbf{z}^{Time}$ (see Fig. 4 (b)(c)). In our FreqMLP, we utilize DCT and IDCT before and after the vanilla MLP. The intuition behind this approach is: High-frequency noise is filtered out of frequency domain features with a low-pass filter, but detailed human motion features (*e.g.*, fast local motions) may also be lost as noise. To address this issue, FreqMLP acts as a trainable frequency-domain filter, allowing us to adaptively adjust the weight of each frequency component in the embedding of 2D joint coordinates (*i.e.*, time-domain features), being a complement to frequency features. These modules are formulated as:

$$\mathbf{z}_k^{'} = \mathrm{MSA}(\mathbf{z}_k), \tag{2}$$

$$\mathbf{z}_k^{Time}, \mathbf{z}_k^{Freq} = \mathbf{z}_k^{'}[: F'], \mathbf{z}_k^{'}[F' :], \tag{3}$$

$$\mathbf{z}_{k+1} = \mathrm{Concat}(\mathrm{FreqMLP}(\mathbf{z}_k^{Time}), \mathrm{MLP}(\mathbf{z}_k^{Freq})), \tag{4}$$

where MSA denotes Multi-head Self-Attention and F' is the number of sampled central frames. The effectiveness of the aforementioned modifications is verified in Sec. 4.4. It's important to recognize that the concatenation operation results in a higher number of tokens for the transformer. However, by restricting the spatial encoder to only observe a limited number of central frames and incorporating a small percentage of low-frequency DCT coefficients to expand its receptive field, we can decrease the overall computation in a flexible manner. This approach not only reduces computational costs but also enhances the model's resistance to noise compared to PoseFormerV1.

Regression head and loss function. Following PoseFormerV1, we use the 1D convolution layer to gather temporal information and a linear projection to obtain the final 3D pose $\mathbf{y} \in \mathbb{R}^{1\times(J\cdot3)}$ for the central frame of the sequence. We use the standard MPJPE (Mean Per Joint Position Error) loss as PoseFormerV1 to train our model.

4. Experiments

4.1. Datasets and Evaluation Metrics

We conduct experiments on two 3D human pose estimation datasets, *i.e.*, Human3.6M [12] and MPI-INF-3DHP [21] to demonstrate the effectiveness of our method. More detailed descriptions of both datasets and their respective evaluation metrics are in the supplementary.

4.2. Implementation Details and Analysis

The proposed method includes three important hyper-parameters that are specific to experimental settings. These include the number of frames (f) used as input in the spatial encoder, the length of the entire input sequence (F) representing the enlarged receptive field, and the number of kept DCT coefficients (n) utilized to incorporate long-range temporal information. If not specified, we simply set $n = f$ for convenience. In practice, they can be further tuned for a flexible speed-accuracy trade-off. When f equals 1, n is set to 3 because a single DCT coefficient may be insufficient to encode temporal information from lengthy input sequences. As f and n are fixed, the computational complexity of the model is predetermined (*i.e.*, the token number for the spatial encoder and that for the feature-fusion module are fixed). We may vary F to effectively expand the model's receptive field from a limited f to an arbitrary value, bringing no additional computational overhead. This enables us to efficiently use long sequences to improve accuracy. We provide details of the hyper-parameters for model architecture and training in the supplementary.

4.3. Comparisons with State-of-the-art Methods

Table 2. Quantitative comparisons with previous transformer-based methods on Human3.6M (in mm). f: number of frames as input to the model, Seq. Len.: length of the entire input sequence (*i.e.*, the effective Receptive Field). The best scores are marked in bold. (*) indicates using an additional pre-training stage and (†) indicates our re-implementation.

Method	f	Seq. Len.	MFLOPs	MPJPE ↓ / P-MPJPE ↓
PoseFormerV1 [41] ICCV'21	27	27	542.1	47.0/–
StridedTrans. [14] TMM'22	81	81	342.5	47.5/–
MixSTE [40](†) CVPR'22	3	3	3420	49.6/38.9
MHFormer [15] CVPR'22	9	9	342.9	47.8/–
MHFormer [15] CVPR'22	27	27	1031.8	45.9/–
P-STMO [29](*) ECCV'22	27	81	163	46.8/–
P-STMO [29](*) ECCV'22	81	81	493	45.6/–
Einfalt *et al.* [9] WACV'23	9	81	543	47.9/–
PoseFormerV2	1	9	**77.2**	49.9/38.7
PoseFormerV2	1	27	**77.2**	48.7/37.8
PoseFormerV2	1	81	**77.2**	**47.6/37.3**
PoseFormerV2	3	9	117.3	49.5/38.5
PoseFormerV2	3	27	117.3	47.9/37.4
PoseFormerV2	3	81	117.3	**47.1/37.3**
PoseFormerV2	9	27	351.7	47.6/37.1
PoseFormerV2	9	81	351.7	**46.0/36.1**
PoseFormerV2	27	243	1054.8	**45.2/35.6**

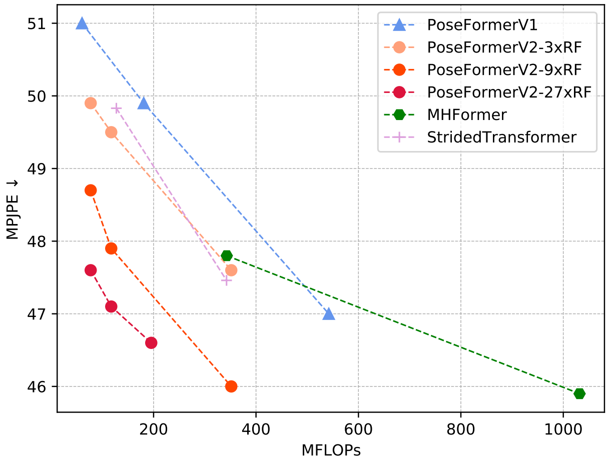

Figure 5. Comparisons of PoseFormerV2 and other state-of-the-art transformer-based methods on Human3.6M (in mm). RF: Receptive Field and $k\times$RF indicates that the RF of PoseFormerV2 is expanded by $k\times$ with a few low-frequency DCT coefficients of the full sequence. The proposed approach outperforms others in terms of speed-accuracy trade-off, and the larger k, the larger improvements over other methods. (Best viewed in color)

Human3.6M. We compare our method with PoseFormerV1 and other transformer-based methods on Human3.6M (Table 2). We demonstrate the flexibility of our model by varying the value of f and the sequence length. Our method is particularly efficient when the expanding ratio (*i.e.*, the ratio of full sequence length to f) is large. For example, with an expanding ratio of 81, it achieves 47.6mm MPJPE with only 77.2 MFLOPs as compared to the 47.8mm MPJPE of MHFormer [15] with 342.9 MFLOPs (4.4× slower). Moreover, with a similar computational budget (around 350 MFLOPs) and the same full sequence length (81), our method achieves 46.0mm MPJPE whereas StridedTransformer [14] obtains 47.5mm MPJPE (3.2%↑). Fig. 5 presents a clearer comparison, showing that

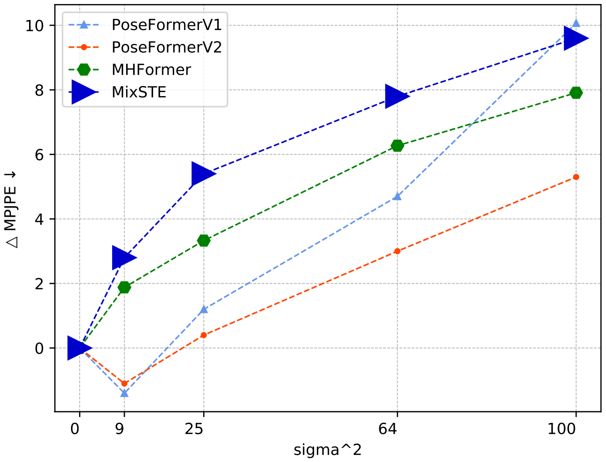

Figure 6. Comparisons of PoseFormerV2 and other transformer-based methods [15, 40, 41] in terms of robustness to noise on Human3.6M. Zero-mean Gaussian noise of standard deviation $sigma$ is added to ground truth 2D detection, and we show their performance drop (△MPJPE, in mm) as $sigma$ increases. The size of markers indicates the computational cost of models.

Table 3. Quantitative comparisons with previous methods on MPI-INF-3DHP. T: the entire sequence length, 1 by default. The best scores are marked in bold. (*) indicates using an additional pre-training stage and (†) indicates our re-implementation.

Method		PCK ↑	AUC ↑	MPJPE ↓
Mehta *et al.* [21]	3DV'17	75.7	39.3	117.6
Mehta *et al.* [22]	ACM ToG'17	76.6	40.4	124.7
Pavllo *et al.* [26] (T=81)	CVPR'19	86.0	51.9	84.0
Pavllo *et al.* [26] (T=243)	CVPR'19	85.5	51.5	84.8
Lin *et al.* [16] (T=25)	BMVC'19	83.6	51.4	79.8
Li *et al.* [13]	CVPR'20	81.2	46.1	99.7
Chen *et al.* [5] (T=81)	TCSVT'21	87.9	54.0	78.8
PoseFormerV1 [41] (T=9)(†)	ICCV'21	95.4	63.2	57.7
MHFormer [15] (T=9)	CVPR'22	93.8	63.3	58.0
MixSTE [40] (T=27)	CVPR'22	94.4	66.5	54.9
P-STMO [29] (T=81)(*)	ECCV'22	**97.9**	75.8	32.2
PoseFormerV2 (T=81)		**97.9**	**78.8**	**27.8**

the proposed method outperforms other transformer-based methods in terms of speed-accuracy trade-off. Note that the methods with an additional pre-training stage and computationally heavy MixSTE [40] (3420 MFLOPs for only 3-frame input) are not included. The improvements of PoseFormerV2 over PoseFormerV1 are provided in Fig. 1.

In order to demonstrate that the inclusion of low-frequency DCT coefficients helps improve the robustness of the proposed method, we make the lifting-based pose estimation task more challenging by adding zero-mean Gaussian noise to the ground-truth 2D detection on the Human3.6M dataset [12] (Fig. 6). To ensure a fair comparison, we keep the input sequence length the same for all methods (in this case, 27 frames). For our method, $f = n = 3$. The experimental evidence reveals that PoseFormerV2 suffers from less performance drop as the standard deviation of Gaussian noise ($sigma$) increases while being more efficient. We observe that the performance of PoseFormerV1 drops drastically as $sigma$ increases from 8 to 10. In contrast, the proposed method presents a more stable trend. Moreover, our method even outperforms MHFormer [15] that incorporates the uncertainty of 2D detectors into the model design. Intriguingly, we find that minor noise may improve the accuracy of 3D pose estimation ($sigma$ = 3).

Table 4. Ablation study on several modifications to PoseFormerV1. We show how a 9-frame PoseFormerV1 is converted to PoseFormerV2 (with 9 DCT coefficients from an 81-frame sequence) step by step. The evaluation is performed on Human3.6M (Protocol 1, in mm). RF indicates Receptive Field.

Step	Description	RF	MPJPE ↓
(0)	Original 9-frame PoseFormerV1.	9	49.9
(1)	Frames are sampled from a longer sequence.	9	49.9
(2)	Append the embedding of DCT coefficients.	81	47.1 (2.8↓)
(3)	Replace the vanilla MLP with FreqMLP.	81	46.0 (3.9↓)

Table 5. Ablation study on the number of frames and the number of DCT coefficients that are used as input to PoseFormerV2. The evaluation is performed on Human3.6M (Protocol 1, in mm).

Frame Number (f)	Coefficient Number (n)	Full Length	MFLOPs	MPJPE
1	1	27	39.2	51.1
1	3	27	77.2	48.7 (2.4↓)
3	1	27	79.4	50.1 (1.0↓)
3	3	27	117.3	47.9 (3.2↓)
9	9	27	351.7	47.6 (3.5↓)

MPI-INF-3DHP. We also compare our method with others on MPI-INF-3DHP [21] (Table 3). We use 9 central frames and the first 9 DCT coefficients from the input 81-frame sequence. The proposed method outperforms other approaches including P-STMO [29] with masked joint pre-training. This result verifies the effectiveness of our method. Our implementation follows [29].

Qualitative comparisons. We provide qualitative comparisons of our method with competitive MHFormer [15] and PoseFormerV1 [41] in Fig. 7. All methods use 81-frame 2D joint sequences as input. To further illustrate the robustness of our approach, we make the pose estimation task more difficult by adding Gaussian noise to the sequential 2D detection of a randomly selected joint (*e.g.*, "left wrist", "right foot"). The proposed method obtains reliable 3D human pose even under highly-deviated 2D detection (indicated by the light-yellow arrows). Note that our model is ∼9× more efficient than MHFormer (3.12 GFLOPs *vs*. 0.35 GFLOPs) and ∼4× more efficient compared to PoseFormerV1 (1.36 GFLOPs *vs*. 0.35 GFLOPs).

4.4. Ablation Study

In this section, we show how a few modifications to PoseFormerV1 bring significant improvements in a step-by-step way. Moreover, to investigate more insights into the frequency-domain representation of input sequences, we reveal the impact of the number of input frames and that of kept DCT coefficients on our method.

Convert PoseFormerV1 into PoseFormerV2. We inherit the overall spatial-temporal architecture from PoseFormerV1 and introduce restrained modifications to its tem-

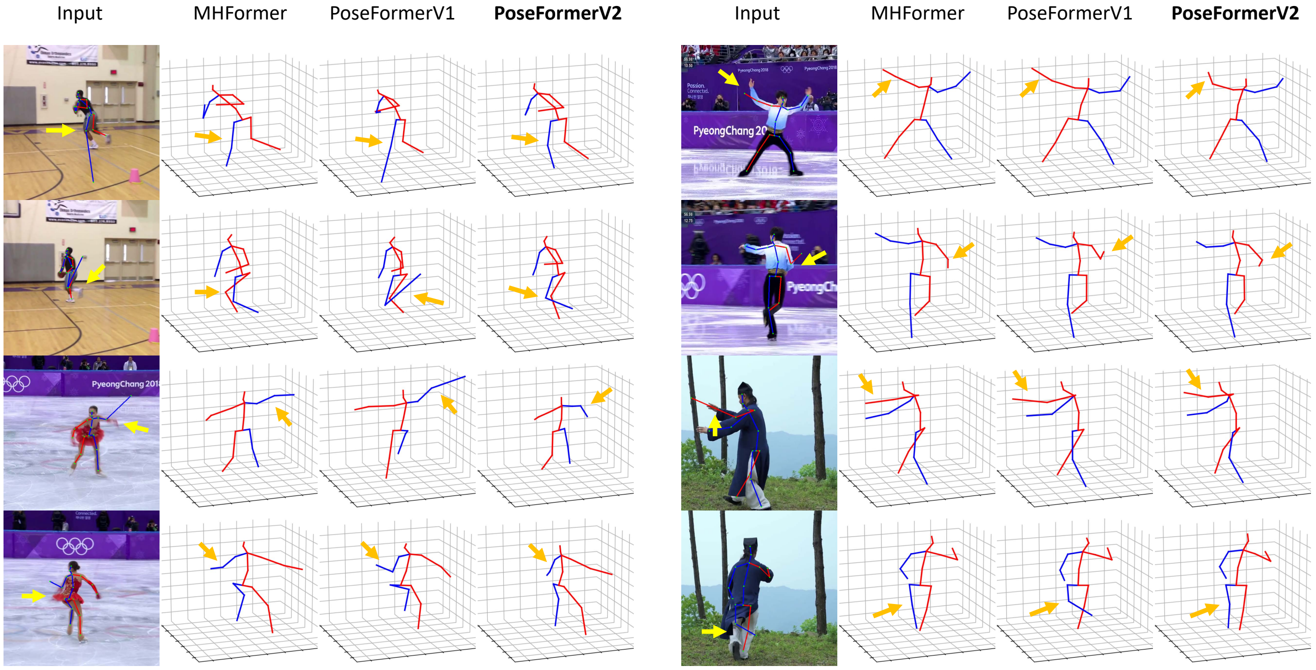

Figure 7. Qualitative comparisons of PoseFormerV2 with MHFormer [15] and PoseFormerV1 [41]. We randomly add Gaussian noise to the 2D detection of a specific joint. We highlight the deviated 2D detection with light-yellow arrows and corresponding 3D pose estimations with orange arrows. PoseFormerV2 shows better robustness to highly noisy input than existing methods.

poral transformer for better multi-domain feature fusion. To exemplify, we illustrate how a 9-frame PoseFormerV1 is converted to PoseFormerV2 step by step: **(1)** the input (*i.e.*, 9 frames) is sampled from a longer sequence (*e.g.*, 81 frames) at the sequence center. This step brings *no* performance improvement or increase in the receptive field since the input to the model is in fact unchanged. **(2)** The output of the spatial encoder of PoseFormerV1, $\mathbf{z}^{Time}$, is appended to the embedding of the first n DCT coefficients (denoted by $\mathbf{z}^{Freq}$) of the complete sequence (81 frames in this case) as input into the temporal encoder. For convenience, we set n to 9. **(3)** We replace the vanilla MLP for $\mathbf{z}^{Time}$ ($\mathbf{z}^{Time}$ and $\mathbf{z}^{Freq}$ already use separate vanilla MLPs before replacement) in the temporal encoder with FreqMLP (details in Sec. 3.2.2). PoseFormerV1 is converted to PoseFormerV2 after these steps, with an enlarged receptive field (from 9 to 81). We present the improvement brought by each step in Table 4. It is worth noting that by introducing 9 DCT coefficients from a longer sequence (*i.e.*, 81 frames), the MPJPE of 9-frame PoseFormerV1 is reduced by **7.8%** (49.9mm *vs.* 46.0mm), which verifies the effectiveness of the proposed DCT representation of input joint sequences.

Number of input frames and DCT coefficients. In Table 5, we investigate the impact of the number of frames (f) as input to the spatial encoder and the number of retained DCT coefficients (n). Here we keep the length of the entire joint sequence fixed, *i.e.*, 27. The baseline model uses only one central frame and one DCT coefficient (f = n = 1). Increasing both parameters brings consistent improvements, and the increase in n translates to more error reduction (2.4↓ for n = 3 *vs.* 1.0↓ for f = 3) since only a few DCT coefficients help capture the global characteristics of the entire sequence. We empirically find that the matched f and n with an expanding ratio of 9 (*i.e.*, f = n = 3) achieve a satisfactory speed-accuracy trade-off.

4.5. Generalization Ability

The proposed frequency-domain approach can generalize to other methods, *e.g.*, MixSTE [40] and MHFormer [15], as they also use transformers for temporal modeling. We improve both methods by incorporating low-frequency DCT coefficients. Details are in supplementary.

5. Conclusion

We present a solution to reconcile two seemingly unrelated or even contracted issues in lifting-based 3D human pose estimation – the efficiency of processing long-sequence input and the robustness to noisy joint detection – simultaneously from a barely explored frequency-domain perspective. The proposed method, PoseFormerV2, exploits a compact frequency representation of long 2D joint sequences to efficiently enlarge the receptive field of the model while improving its robustness. Experimental results show that our method outperforms previous transformer-based methods on Human3.6M and MPI-INF-3DHP.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6836–6846, October 2021. 1

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021. 1

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1

[5] Tianlang Chen, Chen Fang, Xiaohui Shen, Yiheng Zhu, Zhili Chen, and Jiebo Luo. Anatomy-aware 3d human pose estimation with bone-based pose decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 1, 7

[6] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, 2018. 1, 2, 4

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 1, 4

[9] Moritz Einfalt, Katja Ludwig, and Rainer Lienhart. Uplift and upsample: Efficient 3d human pose estimation with uplifting transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2023. 2, 3, 6

[10] John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Adaptive fourier neural operators: Efficient token mixers for transformers. *arXiv preprint arXiv:2111.13587*, 2021. 3

[11] 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. 3

[12] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE TPAMI*, 2014. 1, 2, 3, 4, 6, 7

[13] Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, Chi-Keung Tang, and Kwang-Ting Cheng. Cascaded deep monocular 3d human pose estimation with evolutionary training data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 7

[14] Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu, Pichao Wang, and Wenming Yang. Exploiting temporal contexts with strided transformer for 3d human pose estimation. *IEEE Transactions on Multimedia*, 2022. 2, 3, 6

[15] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13147–13156, June 2022. 2, 3, 6, 7, 8

[16] Jiahao Lin and Gim Hee Lee. Trajectory space factorization for deep video-based 3d human pose estimation. In *BMVC*, 2019. 7

[17] Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-ching Cheung, and Vijayan Asari. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In *CVPR*, 2020. 1

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1

[19] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*, pages 474–489. Springer, 2020. 2, 3

[20] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9489–9497, 2019. 2, 3

[21] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017. 3, 6, 7

[22] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017. 7

[23] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *ECCV*, 2020. 1

[24] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 1

[25] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3D human pose estimation. In *CVPR*, 2018. 1

[26] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, 2019. 7

[27] William B Pennebaker and Joan L Mitchell. *JPEG: Still image data compression standard*. Springer Science & Business Media, 1992. 3

[28] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *Advances in Neural Information Processing Systems*, 34:980–993, 2021. 3

[29] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. *arXiv preprint arXiv:2203.07628*, 2022. 2, 3, 6, 7

[30] Athanassios Skodras, Charilaos Christopoulos, and Touradj Ebrahimi. The jpeg 2000 still image compression standard. *IEEE Signal processing magazine*, 18(5):36–58, 2001. 3

[31] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020. 1

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017. 1

[33] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3d pose estimation from videos. In *European Conference on Computer Vision*, pages 764–780. Springer, 2020. 1

[34] Zhenyu Wang, Hao Luo, Pichao WANG, Feng Ding, Fan Wang, and Hao Li. VTC-LFC: Vision transformer compression with low-frequency components. In *Thirty-Sixth Conference on Neural Information Processing Systems*, 2022. 2, 3

[35] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022. 3

[36] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 3

[37] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 3

[38] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3333–3343, 2022. 1

[39] Ailing Zeng, Xiao Sun, Fuyang Huang, Minhao Liu, Qiang Xu, and Stephen Lin. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In *ECCV*, 2020. 1

[40] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13232–13242, 2022. 2, 3, 6, 7, 8

[41] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11656–11665, October 2021. 1, 2, 3, 4, 6, 7, 8

[42] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1