MPM: A Unified 2D-3D Human Pose Representation via Masked Pose Modeling

Zhenyu Zhang¹ * Wenhao Chai² * Zhongyu Jiang² Tian Ye³
Mingli Song¹ Jenq-Neng Hwang² Gaoang Wang^{4†}
Zhejiang University¹ University of Washington²
Jimei University³ ZJU-UIUC Institute, Zhejiang University⁴

{zhenyuzhang, brooksong}@zju.edu.cn, {wchai, zyjiang, hwang}@uw.edu, 201921114031@jmu.edu.cn, gaoangwang@intl.zju.edu.cn

Abstract

Estimating 3D human poses only from a 2D human pose sequence is thoroughly explored in recent years. Yet, prior to this, no such work has attempted to unify 2D and 3D pose representations in the shared feature space. In this paper, we propose MPM, a unified 2D-3D human pose representation framework via masked pose modeling. We treat 2D and 3D poses as two different modalities like vision and language and build a single-stream transformerbased architecture. We apply three pretext tasks, which are masked 2D pose modeling, masked 3D pose modeling, and masked 2D pose lifting to pre-train our network and use full-supervision to perform further fine-tuning. A high masking ratio of 72.5 % in total with a spatio-temporal mask sampling strategy leading to better relation modeling both in spatial and temporal domains. MPM can handle multiple tasks including 3D human pose estimation, 3D pose estimation from occluded 2D pose, and 3D pose completion in a single framework. We conduct extensive experiments and ablation studies on several widely used human pose datasets and achieve state-of-the-art performance on Human3.6M and MPI-INF-3DHP. Codes and model checkpoints are available at this https URL.

1. Introduction

3D human pose estimation (HPE) tasks aim to estimate 3D human joints position from a single image or a video clip. 3D HPE has a wide range of applications in various computer vision tasks, such as action recognition [22, 38, 49], multiple object tracking [15], human parsing [51], and other human-centric perception tasks. Many 3D human pose estimation methods are based on the immediate results of 2D human pose estimation [4, 6], the so-called lifting paradigm. Lifting methods fully utilize the image/video-2D pose datasets (e.g., COCO [28]) and 2D-

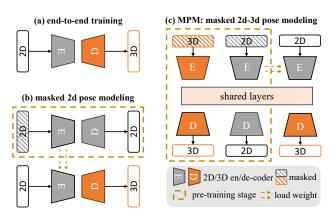


Figure 1. Comparison of the existing 3d human pose estimation lifting method and our *MPM*. (a) End-to-end training without pretraining using any backbone; (b) Pre-training with mask 2D pose modeling and then fine-tuning [40]; (c) our *MPM*: pre-trained with both mask 2D and 3D pose modeling, in which most of the parameters in shared layers to learn a unified representation.

3D pose datasets (*e.g.*, Human3.6M [18] to form the two-stage methods. Yet, existing 2D-3D pose datasets either lack enough diversity in laboratorial environments [18] or lack adequate quantity and accuracy in the wild [34], resulting in a large domain gap among different datasets and also in real-world scenarios. There are few prior arts aim to improve the generalization [14] or adaptation [2,13] capability among various domains. Nonetheless, developing a human pose pre-training paradigm by utilizing all the 2D-3D pose datasets is far from fully explored.

As for the lifting network architectures, early works are based on fully-connected layers networks [33], 3D convolution networks [36], and graph neural networks [48]. Recently, transformer-based [44] methods also have achieved state-of-the-art performance in 3D human pose estimation tasks [25, 53, 54] following lifting paradigm. Empirically, the amount of data used to train the model with transformer architecture is often much more than those with other back-

^{*} Equal contribution. † Corresponding author.

bones. On the other hands, the pre-training paradigm in transformers are widely discussed, which might also benefit human pose estimation tasks. Inspired by some previously proposed self-supervised pre-training method like masked modeling [16, 40] and also the transformer architecture design in multi-modality area [3,31], we propose *MPM*, a unified 2D-3D human pose representation network pre-trained by masked pose modeling paradigm.

We treat 2D and 3D pose as two different modalities like vision and language and build a single-stream transformer-based architecture. To be specific, as shown in Figure 1, we first use two separate encoders to embed 2D and 3D poses into a shared embedding space. After that, the 2D or 3D pose embedding are put into shared transformer layers, which contains most of the parameters of our network. Note that only one of the 2D and 3D poses are put in at once. Finally, two separate decoders are used to decode 2D and 3D poses from embedding space.

We use two training stages including masked modeling pre-training (stage I) and full-supervised fine-tuning (stage II) to train our network. In stage I, we apply three pretext tasks including (1) masked 2D pose modeling, (2) masked 3D pose modeling, and (3) masked 2D pose lifting. Note that we conduct masked operation both in spatial and temporal direction. After pre-training on those masked modeling pretext tasks, the model have learned the prior knowledge of spatial and temporal relations of both 2D and 3D human pose and the lifting information between them. In stage II, we use 2D-3D pose pairs to further perform fine-tuning since the network is trained with masked pose input, which cause the gap when unmasked poses are set as input.

Our contributions are summarized as follows:

- We are the first to treat 2D and 3D pose as two different modalities in a shared embedding space and build a single-stream transformer-based architecture for 3D human pose estimation and pose completion tasks.
- We apply three masked modeling based pretext tasks for human pose pre-training to learn spatial and temporal relations.
- We conduct extensive experiments and ablation studies on multiple widely used human pose datasets and achieve state-of-the-art performance on MPI-INF-3DHP and Human3.6M benchmarks.

2. Related Works

2.1. 3D Human Pose Estimation in Video

Estimating 3D human pose only from 2D human pose sequences is the common paradigm for 3D human pose estimation in video and has been thoroughly explored in recent years. Pavllo *et al.* [36] leverage dilated temporal convolutions with semi-supervised way to improve

3D pose estimation in videos. Martinez *et al.* [33] propose an MLP-based network on lifting 2D to 3D poses. Transformer-based networks achieve state-of-the-art performance recently. Zheng *et al.* [54] is the first to introduce a transformer into 3D human pose estimation task. Li *et al.* [25] propose a multi-hypothesis transformer that learns spatio-temporal representations of multiple plausible pose hypotheses. Zhao *et al.* [53] aim to alleviate the negative effect from the length of the input joint sequences and the quality of detected 2D joints in the human pose transformer. Those prior arts show the effectiveness of transformer architectures for 3D human pose estimation task in video. We take the successful experiences to design our network.

2.2. Transformer-based Multimodal Architecture

As we treat 2D and 3D poses as two different modalities like vision and language, we also learn from the transformer-based architecture from multimodal [3, 47] field. According to the network structures, multimodal transformer can be divided into single-stream (*e.g.*, Uniter [7], Visualbert [21], VI-BERT [42]), multi-stream (*e.g.*, ViLBERT [30], Lxmert [43], ActBERT [55]), hybrid-stream (*e.g.*, InterBERT [27]), *etc.* To learn a unified 2D/3D human pose representation, we follow the two-stream architecture design, in which a large amount of the parameters are shared between 2D and 3D poses. The encoders and decoders are relatively light compared to the shared transformer layers. In this case, we hope the prior knowledge can be maximized between 2D and 3D human poses in spatial and temporal domains as well as the 2D-3D lifting priors.

2.3. Network Pre-training via Masked Modeling

Masked modeling as the pre-training task is first used in natural language processing (NLP) field. BERT [9] masks out a portion of the input language tokens and train models to predict the missing content. In computer vision (CV) field, Beit [37] applies masked image modeling with a vector-quantized visual tokenizer. MAE [16] and Video-MAE [12] also show the effectiveness of mask modeling by masking random patches of the input image or video and reconstruct the missing pixels. P-STMO [40] is the first to introduce masked pose modeling in the 3D human pose estimation task. It randomly masks some of the 2D poses, and reconstructs the complete sequences with a model consisting of spatial encoder, temporal encoder, and decoder. Yet, P-STMO has not explored the masked 3D pose modeling as well as the unified representation of 2D and 3D human pose. Besides, pre-training with multiple 2D-3D pose datasets are not explored either. In this paper, we show that 3D poses can also be exploited in pre-training tasks using the same way as 2D and even further in a unified manner.

Stage I: Pre-training via Masked Pose Modeling

Stage II: Fine-tuning

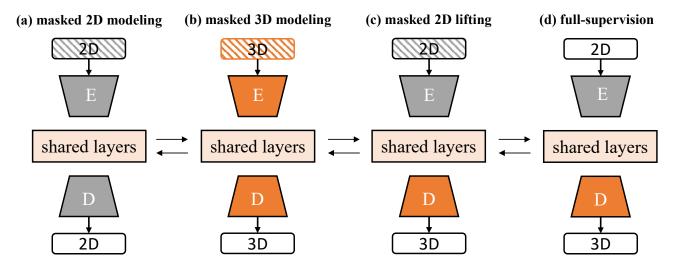


Figure 2. We conduct two training stages to train the proposed network. In stage I, we apply three pretext tasks including (a) masked 2D modeling, (b) masked 3D modeling, and (c) masked 2D lifting. After pre-training on those masked modeling pretext tasks, the model has learned the prior knowledge of spatial and temporal relations of both 2D and 3D human poses and the lifting information between them. In stage II, we use unmasked 2D-3D pose pairs to further fine-tune the network.

3. Methodology

3.1. Architecture

2D/3D encoder. We use a simple MLP as the architecture for both the 2D and 3D pose encoders. We use 1D convolution with a kernel size of 1 as a fully connected layer. 2D and 3D poses are embedded into the same-dimension features. Note that we encode the input pose sequence *frame-by-frame*, which has no temporal relation modeling. The encoders are relatively light as shown in Table 7 compared to shared transformer layers since we aim to force 2D and 3D poses to share the unified representation. This process can be formulated as:

$$f_{2D}^0 = \mathcal{E}_{2D}(P_{2D}^{in}), \quad f_{3D}^0 = \mathcal{E}_{3D}(P_{3D}^{in}),$$
 (1)

where $\mathbf{P}_{2D}^{in} \in \mathbb{R}^{l \times J \times 2J}$ and \mathbf{P}_{3D}^{in} are the input 2D and 3D pose sequences, $\mathcal{E}_{2D}(\cdot)$ and $\mathcal{E}_{3D}(\cdot)$ are 2D and 3D encoders, and \mathbf{f}_{2D}^0 and \mathbf{f}_{3D}^0 are 2D and 3D pose embeddings before being put in shared layers which share the common space.

Shared layers. In shared layers, we use both MLP and transformer-based architectures to process the information on pose sequences. Shared layers contain the most parameters of our network.

We model spatial relations through a shared MLP, which is much heavier than the 2D/3D encoders, to model temporal relations through shared temporal transformer layers as shown in Figure 3.

The pose embedding input is first calculated by the MLP-based spatial encoder. This process can be formulated as:

$$f_{2D}^n = \mathcal{S}(f_{2D}^0), \quad f_{3D}^n = \mathcal{S}(f_{2D}^0),$$
 (2)

where $S(\cdot)$ is the shared spatial layers.

The transformer blocks in shared layers follow the scaled dot-product attention [44]. The attention computing of query key, and value matrices Q, K, V in each head are formulated by:

$$Attention(Q, K, V) = SoftMax(\frac{QK^{T}}{\sqrt{d_m}})V, \quad (3)$$

where $Q,K,V\in\mathbb{R}^{N\times d_m}, N$ indicates the number of tokens and d_m indicates the dimension of each token. In our framework, N is the pose sequence length, and d_m is the feature dimension of the embedding.

The shared temporal layers is based on a vanilla transformer architecture. We treat frames as tokens to model the temporal relation of poses in the same sequence. The process can be formulated as:

$$f_{2D}^l = \mathcal{T}(f_{2D}^n), \quad f_{3D}^l = \mathcal{T}(f_{3D}^n),$$
 (4)

where $\mathcal{T}(\cdot)$ is the shared temporal layers. $f_{2D}^n \in \mathbb{R}^{L \times C}$,

2D/3D decoder. As for the 2D decoder, we use a single transformer block including the self-attention layer and

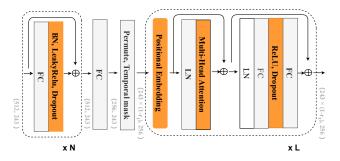


Figure 3. Detailed architecture of the shared transformer layers. We take 243 frames as the tensor size in the pipeline.

the feed forward networks (FFNs), aiming to encourage the shared transformer layers to learn better universal features.

The 3D decoder shares the same architecture with the 2D decoder in stage I. However, we found that it is difficult to regress 3D human pose through a light decoder without performance drop in stage II. Therefore, we additionally use stride-transformer layers [24] to enhance the lifting ability of the 3D decoder. To be specific, the FFN layer is replaced by the convolution layer. To utilize the pre-trained model, the weights of the 3D decoder in stage I are loaded into the first block as the initialization in stage II fine-tuning. This process can be formulated as:

$$P_{2D}^{out} = \mathcal{D}_{2D}(f_{2D}^l), \quad P_{3D}^{out} = \mathcal{D}_{3D}(f_{3D}^l),$$
 (5)

where P_{2D}^{out} and P_{3D}^{out} are the output 2D and 3D pose sequence, $\mathcal{D}_{2D}(\cdot)$ and $\mathcal{D}_{3D}(\cdot)$ are 2D and 3D encoders, and f_{2D}^l and f_{3D}^l are 2D and 3D pose embedding after shared transformer layers which share the common space.

3.2. Stage I: Pre-training via Masked Pose Modeling

3.2.1 Masked sampling strategies

We conduct randomly masked sampling frame-by-frame as shown in Figure 7. In this case, the masked joints between adjacent frames could not be exactly the same. We claim this kind of masked sampling strategy can help to learn both spatial and temporal relations, as evidenced from ablation studies among different masked sampling strategies in Section 4.4.3. The masked joints are replaced by a shared learnable vector v^S before the encoder. It is also padded by a shared learnable vector v^T before the decoder, which is similar to [16].

3.2.2 Pretext tasks

Masked 2D/3D pose modeling. We use the L2 loss between the masked and reconstructed 2D/3D pose sequences:

$$P_i^{in} = M_i, \quad P_i^{out} = (\mathcal{E}_i \circ \mathcal{S} \circ \mathcal{T} \circ \mathcal{D}_i)(M_i), \quad (6)$$

for $i \in \{2D, 3D\}$ where M_i is the masked pose sequence. Then the masked pose modeling loss can be calculated by:

$$\mathcal{L}_{m2m} = \|P_{2D}^{out} - P_{2D}^{in}\|_2, \quad \mathcal{L}_{m3m} = \|P_{3D}^{out} - P_{3D}^{in}\|_2.$$
 (7)

Masked 2D pose lifting. We also use L2 loss to calculate the loss function of masked 2D pose lifting:

$$P_{2D}^{in} = M_{2D}, \quad P_{3D}^{out} = (\mathcal{E}_{2D} \circ \mathcal{S} \circ \mathcal{T} \circ \mathcal{D}_{3D})(M_{2D}),$$
 (8)

where M_{2D} is the masked 2D pose sequence. Then the masked pose modeling loss can be calculated by:

$$\mathcal{L}_{m2l} = \|P_{3D}^{gt} - P_{3D}^{out}\|_2, \tag{9}$$

where P_{3D}^{gt} is the ground truth 3D pose sequence for supervision.

Stage I conclusion. We conduct those three pre-training tasks iteratively with the same frequency. The overall loss function in stage I is:

$$\mathcal{L}_{\text{stage I}} = \lambda_{\text{m2m}} \mathcal{L}_{\text{m2m}} + \lambda_{\text{m3m}} \mathcal{L}_{\text{m3m}} + \lambda_{\text{m2l}} \mathcal{L}_{\text{m2l}}, \quad (10)$$

where λ_{m2m} , λ_{m3m} , and λ_{m2l} are weight factors.

3.3. Stage II: Fine-tuning via Full-supervision

In stage II, we use 2D-3D pose pairs to further perform fine-tuning since the network is trained with masked pose input, which causes the input gap when unmasked poses are set as input. We input a 2D sequence to the model and get the 3D pose of the middle frame following [36] as

$$P_{3D}^{out} = (\mathcal{E}_{2D} \circ \mathcal{S} \circ \mathcal{T} \circ \mathcal{D}_{3D})(P_{2D}). \tag{11}$$

Here we still use L2 loss between the prediction and ground truth 3D pose formulated as:

$$\mathcal{L}_{ft} = \|PM_{3D}^{out} - PM_{3D}^{gt}\|_2, \tag{12}$$

where PM_{3D}^{out} and PM_{3D}^{gt} represent the single pose in the middle frame of the pose sequence and the corresponding ground truth, respectively.

We also utilize a simple linear projection after the shared layer to obtain the whole 3D pose sequence as extra supervision. The loss function denotes as:

$$\mathcal{L}_{\text{ft-seq}} = \|P_{out}^{3D} - P_{gt}^{3D}\|_2.$$
 (13)

Stage II conclusion. The overall loss function in stage II is:

$$\mathcal{L}_{\text{stage II}} = \lambda_{\text{ft}} \mathcal{L}_{\text{ft}} + \lambda_{\text{ft-seq}} \mathcal{L}_{\text{ft-seq}}, \tag{14}$$

where $\lambda_{\rm ft}$ and $\lambda_{\rm ft\text{-}seq}$ are weight factors.

Protocol #1 (GT)	Venue	# Frames	Dir.	Disc.	Eat.	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Somke	Wait	WalkD.	Walk	WalkT.	Average
Lin et al. [26]	BMVC'18	50	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	32.8
Liu et al. [29]	CVPR'20	243	34.5	37.1	33.6	34.2	32.9	37.1	39.6	35.8	40.7	41.4	33.0	33.8	33.0	26.6	26.9	34.7
Zeng et al. [50]	ECCV'20	243	34.8	32.1	28.5	30.7	31.4	36.9	35.6	30.5	38.9	40.5	32.5	31.0	29.9	22.5	24.5	32.0
Chen et al. [5]	TCSVT'21	243	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	32.3
Zheng et al. [54]†	ICCV'21	81	30.0	33.6	29.9	31.0	30.2	33.3	34.8	31.4	37.8	38.6	31.7	31.5	29.0	23.3	23.1	31.3
Shan et al. [41]	MM'21	243	29.5	30.8	28.8	29.1	30.7	35.2	31.7	27.8	34.5	36.0	30.3	29.4	28.9	24.1	24.7	30.1
Shan et al. [40]†	ECCV'22	81	28.5	30.1	28.6	<u>27.9</u>	29.8	<u>33.2</u>	31.3	<u>27.8</u>	36.0	37.4	<u>29.7</u>	29.5	28.1	21.0	21.0	29.3
MPM (ours)	_	81	30.4	32.2	29.2	28.4	29.3	33.7	34.2	29.3	40.0	35.6	30.8	30.7	27.4	20.7	20.7	30.2
MPM (ours)	_	243	26.7	28.1	27.1	26.2	27.8	30.9	30.5	26.0	33.9	33.1	27.3	28.1	24.7	18.3	18.5	27.2
Protocol #1 (CPN)	Venue	# Frames	Dir.	Disc.	Eat.	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Somke	Wait	WalkD.	Walk	WalkT.	Average
	CVPR'19	243	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
Pavllo et al. [36] Lin et al. [26]	BMVC'19	243	42.5	44.8	42.6	45.6	48.1	57.1	52.6	44.3	56.5	64.5	47.1 47.4	44.0	49.0	32.8	35.9	46.8
Liu et al. [29]	CVPR'20	243	42.3	44.8	41.1	44.2	48.3	54.1	43.4	42.2	56.2	63.6	45.3	43.5	45.3	31.3	32.2	45.1
Zeng et al. [50]	ECCV'20	243	46.6	44.8	43.9	44.9	45.8	49.6	46.5	40.0	53.4	61.1	45.3	42.6	43.3	31.5	32.2	44.8
Wang et al. [45]	ECCV 20 ECCV'20	96	41.3	43.9	44.0	42.2	43.8	57.1	42.2	43.2	57.3	61.3	47.0	43.5	47.0	32.6	31.8	45.6
Chen et al. [5]	TCSVT'21	81	42.1	43.9	41.0	43.8	46.1	53.5	42.4	43.2	53.9	60.5	45.7	42.1	46.2	32.0	33.8	44.6
Zheng et al. [54]†	ICCV'21	81	41.5	44.8	39.8	42.5	46.5	51.6	42.4	42.0	53.3	60.7	45.5	43.3	46.1	31.8	32.2	44.3
Chen et al. [5]	TCSVT'21	243	41.4	43.5	40.1	42.9	46.6	51.0	41.7	42.3	53.9	60.2	45.4	41.7	46.0	31.5	32.7	44.1
Hu et al. [17]	MM'21	243	38.0	43.3	39.1	39.4	45.8	53.6	41.4	41.4	55.5	61.9	44.6	41.9	44.5	31.6	29.4	43.4
Li et al. [24]†	TMM'22	351	39.9	43.4	40.0	40.9	46.4	50.6	42.1	39.8	55.8	61.6	44.9	43.3	44.9	29.9	30.3	43.4
Shan et al. [40]†	ECCV'22	243	38.4	42.1	39.8	40.2	45.2	48.9	40.4	38.3	53.8	57.3	43.9	41.6	42.2	29.3	29.3	42.1
MPM (ours)		81	39.7	43.2	39.9	42.7	45.8	52.7	41.7	41.1	55.3		45.1	42.7	43.8	29.4	30.2	
MPM (ours) MPM (ours)	_	243	39.7	43.2 42.1	39.9	40.3	45.8 45.2	50.9	40.1	39.6	53.9	60.5 59.4	45.1 43.0	41.3	43.8 42.2	29.4 28.9	30.2 29.1	43.6 42.3
MFM (ours)		243	39.1	42.1	39.2	40.3	43.2	30.9	40.1	39.0	33.9	39.4	43.0	41.3	42.2	20.7	27.1	42.3
Protocol #2 (CPN)	Venue	# Frames	Dir.	Disc.	Eat.	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Somke	Wait	WalkD.	Walk	WalkT.	Average
Lin et al. [26]	BMVC'19	50	32.5	35.3	34.3	36.2	37.8	43.0	33.0	32.2	45.7	51.8	38.4	32.8	37.5	25.8	28.9	36.8
Pavllo et al. [36]	CVPR'19	243	34.1	36.1	34.4	37.2	36.4	42.2	34.4	33.6	45.0	52.5	37.4	33.8	37.8	25.6	27.3	36.5
Xu et al. [46]	CVPR'20	50	31.0	34.8	34.7	34.4	36.2	43.9	31.6	33.5	42.3	49.0	37.1	33.0	39.1	26.9	31.9	36.2
Liu et al. [29]	CVPR'20	243	32.3	35.2	33.3	35.8	35.9	41.5	33.2	32.7	44.6	50.9	37.0	32.4	37.0	25.2	27.2	35.6
Wang et al. [45]	ECCV'20	96	32.9	35.2	35.6	34.4	36.4	42.7	31.2	32.5	45.6	50.2	37.3	32.8	36.3	26.0	23.9	35.5
Chen et al. [5]	TCSVT'21	81	33.1	35.3	33.4	35.9	36.1	41.7	32.8	33.3	42.6	49.4	37.0	32.7	36.5	25.5	27.9	35.6
Zheng et al. [54]†	ICCV'21	81	32.5	34.8	32.6	34.6	35.3	39.5	32.1	32.0	42.8	48.5	34.8	32.4	35.3	24.5	26.0	34.6
Shan et al. [40]†	ECCV'22	243	31.3	35.2	32.9	33.9	35.4	39.3	32.5	<u>31.5</u>	44.6	48.2	36.3	32.9	34.4	23.8	23.9	34.4
MPM (ours)	-	81	31.9	35.3	33.2	35.0	35.6	40.5	33.0	32.4	44.8	48.7	37.3	33.4	34.4	24.2	25.7	35.0
MPM (ours)	_	243	32.4	34.7	32.3	33.6	35.3	40.4	31.8	31.5	44.3	49.4	<u>35.4</u>	32.2	34.2	23.7	24.3	34.4

Table 1. Quantitative comparison of Mean Per Joint Position Error between the estimated 3D pose and the ground truth 3D pose on Human3.6M under Protocols #1 & #2 using the detected 2D pose as input. Top-table: results under Protocol #1 (MPJPE). Bottom-table: results under Protocol #2 (P-MPJPE). † denotes a Transformer-based model. The best and second scores are marked in bold and underline.

4. Experiments

4.1. Datasets and Metrics

We use three widely used 2D-3D human pose datasets to train our network including Human3.6M [18], MPI-INF-3DHP [34], and AMASS [32]. We evaluate the performance on Human3.6M and MPI-INF-3DHP.

Human3.6M. Human3.6M dataset, which contains 3.6-million frames of corresponding 2D and 3D human poses, is a video and mocap dataset of 5 female and 6 male subjects. According to the previous works [5, 10, 11], we choose 5 subjects (S1, S5, S6, S7, S8) for training, and the other 2 subjects (S9 and S11) for evaluation. We report the Mean Per Joint Position Error (MPJPE) as the metric of Protocol #1 as well as Procrusts analysis MPJPE (P-MPJPE) as the metric of Protocol #2.

MPI-INF-3DHP. Compared to Human3.6M, MPI-INF-3DHP is a more challenging 3D human pose dataset captured in the wild. There are 8 subjects with 8 actions captured by 14 cameras coving a greater diversity of poses. We follow the setting of evaluation in [5] with three common metrics PCK, AUC, and MPJPE.

AMASS. AMASS [32] is a large database of human motion unifying different optical marker-based motion capture datasets. The consistent representation of AMASS makes it readily useful for animation, visualization, and generating training data for deep learning. It totally has more than 40 hours of duration over 300 subjects, and more than 11,000 motions are involved.

4.2. Implementation details

Training settings. We implement our proposed method using PyTorch [35] on Four NVIDIA RTX 3090 GPUs. For both of stage I and stage II, we use Adam [19] optimizer and train for 50 epochs with batch size of 1024. The learning rate is set to $1e^{-3}$ for stage I and $7e^{-4}$ for stage II, and we set a weight decay factor of 0.98 of each epoch. As a video-based method, we choose two different input sequence length of 81 and 243. We also use horizontal flip and PoseAug [14] as the data augmentation approach. The loss weight factors $\lambda_{\rm m2m}$, $\lambda_{\rm m3m}$, $\lambda_{\rm m2l}$, $\lambda_{\rm ft}$ and $\lambda_{\rm ft-seq}$ are set to 5.0, 2.0, 1.0, 2.0, 1.0, respectively.

Datasets mixing settings. Since we use multiple datasets for training, we sample data randomly on each batch. Note that we use 16 keypoints definition among all the datasets following [14].

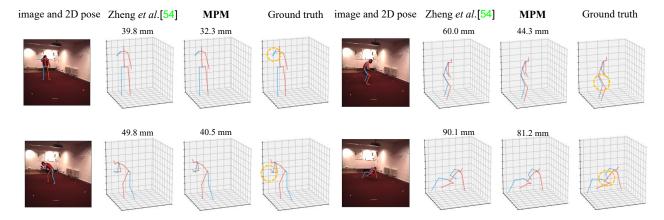


Figure 4. Qualitative visualization comparison with SOTA method [54] and ground truth on Human 3.6M benchmarks.

Method	Venue	# Frames	PCK (†)	AUC (†)	MPJPE (\downarrow)
Mehta et al. [34]	3DV'17	1	75.7	39.3	117.6
Pavllo et al. [36]	CVPR'19	81	86.0	51.9	84.0
Pavllo et al. [36]	CVPR'19	243	85.5	51.5	84.8
Lin et al. [26]	BMVC'19	25	83.6	51.4	79.8
Li et al. [23]	CVPR'20	1	81.2	46.1	99.7
Chen et al. [5]	TCSVT'21	81	87.9	54.0	78.8
Zheng et al. [54]	ICCV'21	81	88.6	56.4	77.1
Zhang et al. [52]	CVPR'22	243	94.4	66.5	54.9
Hu et al. [17]	MM'21	81	97.9	69.5	42.5
Shan et al. [40]	ECCV'22	81	97.9	75.8	32.2
MPM (ours)	_	81	98.6	79.3	29.8
MPM (ours)	_	243	<u>98.4</u>	80.0	29.1

Table 2. Quantitative comparison with previous methods on MPI-INF-3DHP. PCK, AUC, and MPJPE metrics are reported. The best and second scores are marked in bold and underline respecively.

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

4.3.1 3D Human Pose Estimation

Results on Human3.6M. We compare our methods with existing state-of-the-art methods on Human3.6M dataset. We report the results of all the 15 actions in the test subjects (S9 and S11) as shown in Table 1. Following [24,40,54], we use 2D poses detected by the CPN [6] detector. Our method achieves comparable performance.

Results on MPI-INF-3DHP. We report our result and compared it with state-of-the-art methods on MPI-INF-3DHP [1] dataset as shown in Table 2. We only adopt ground truth of 2D poses as input.

Qualitative visualization. Figure 4 demonstrates the qualitative visualization comparison with the previous works on Human3.6M benchmark.

Occ. Body Parts	MPM (ours)	GFPose $(S=1)$ [8]	${\rm GFPose}\;(S=200)$	Li et al. [20]
1 Joint 2 Joints	63.7 69.1	71.7 78.3	37.8 39.6	58.8 64.6
2 Legs	91.8	108.6	53.5	-
2 Arms Left Leg + Left Arm	101.0 84.8	116.9 106.8	60.0 54.6	_
Right Leg + Right Arm	84.9	109.7	53.1	-

Table 3. Recover 3D pose from partial 2D observation under Protocol #1. S denotes sample number in multi-hypothesis setting.

Occ. Body Parts	MPM (ours)	GFPose $(S = 1)$ [8]	GFPose $(S=200)$
Right Leg	15.1	13.0	5.2
Left Leg	14.2	14.3	5.8
Left Arm	23.1	25.5	9.4
Right Arm	22.3	22.4	8.9

Table 4. Recover 3D pose from partial 3D observation under Protocol $\#1.\ S$ denotes sample number in multi-hypothesis setting.

4.3.2 3D Pose Completion

From incomplete 2D pose. In real-world application 2D pose estimation algorithms and MoCap systems often suffer from occlusions, which result in incomplete detected 2D pose. Pre-trained by masked 2D pose lifting, *MPM* can also help to recover an intact 3D human pose from incomplete 2D observations at either the joint level or body part level. As shown in Table 3, our method achieves comparable results even though GFPose [8] uses 200 samples in multi-hypothesis setting and also with camera intrinsic, Li *et al.* [20] train different models to deal with different numbers of missing joints.

From incomplete 3D pose. Fitting to partial 3D observations also has many potential downstream applications, *e.g.*, generating the missing part of body for person in VR. Shown in Table 4, *MPM* can be directly used to recover missing 3D body parts given partial 3D observations since masked 3D pose modeling is applied in pre-training stage.

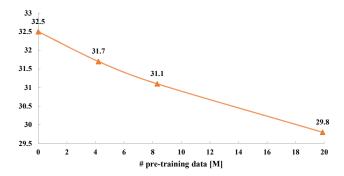


Figure 5. Performance under different pre-training data scales shown in Table 5 under Protocol #1 on MPI-INF-3DHP.

4.4. Ablation Study

In this section, we conduct extensive ablation studies to answer the questions as follows.

4.4.1 Does MPM Follow Data Scaling Law?

We claim that a good pre-training framework could benifit from larger pre-training data. To verify this kind of data scaling law in *MPM*, we pre-train our network under three different training data scales as shown in Table 5. Figure 5 demonstrates the plot of performance under different pre-training data scales. We show that our *MPM* could benefit from large training data in pre-training stage.

Setting	Human3.6M	MPI-INF-3DHP	AMASS	# Samples
Small	√ half	√ _{quarter}		4.21 M
Base	✓	√half		8.32 M
Large	✓	\checkmark	\checkmark	19.88 M

Table 5. Three different training data scales. $\sqrt{}$ half and $\sqrt{}$ quarter denotes using half and quarter of the data.

4.4.2 Does *MPM* Help Learning Human Priors?

In this part, we discuss the potential role of *MPM* in learning human priors and how can *MPM* further benefit 3D human pose estimation task. We design the experiments that only use limited annotated data for full-supervised training. We only use 1/20 S1 and S5 subjects on Human3.6M with 2D pose detected by CPN and evaluate the performance of the models on S9 and S11 subjects on Human3.6M. As shown in Table 6, compared to other end-to-end network, our *MPM* outperforms in this limited data setting, which shows that the pre-training stage actually helps learning human priors. Surprisingly, *MPM* also shows better performance than pre-training method P-STMO, in which only the masked 2D pose modeling task is applied in the pre-training stage. Table 7 demonstrates the number of param-

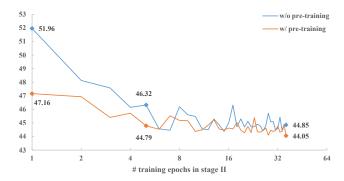


Figure 6. The performance during stage II training process under Protocol #1 on Human3.6M for 40 epochs.

eters in different components. The shared parameters take the majority, which shows unifying 2D and 3D human pose representations leads to more robust and generic human priors.

We also trace the performance during training process for both models with and without pre-training. Figure 6 shows that a longer training schedule gives a noticeable performance improvement. Compared with the model trained from scratch, the pre-training model demonstrates faster convergence speed and better performance.

Method	Venue	# Frames	MPJPE (↓)	PA-MPJPE (↓)
Videopose3d [36]	CVPR'19	243	70.0	49.8
Zhang <i>et al</i> . [52]	CVPR'22	243	74.9	56.6
Zheng et al. [54]	ICCV'21	81	73.1	51.6
Shan et al. [40]	ECCV'22	243	70.2	49.7
Zhao et al. [53]	CVPR'23	81	70.0	49.9
MPM (ours)	_	81	<u>68.7</u>	48.0
MPM (ours)	_	243	67.9	<u>49.2</u>

Table 6. Performance under limited full-supervision training data under Protocol #1 and Protocol #2 on Human3.6M.

# Params. [K]	FLOPs [M]
17.9	8.9
26.1	12.9
2763.0	1339.7
532.2	259.3
534.3	261.3
3941.1	380.6
	17.9 26.1 2763.0 532.2 534.3

Table 7. Analysis on computational complexity of each module.

4.4.3 Does Mask Sampling Strategies Matter?

We conduct ablation studies on three different mask sampling strategies as shown in Figure 7 including: 1) spatial masking or called tube masking, in which certain joints are masked over all the frames; 2) temporal masking, in which all the joints in certain frames are masked; and 3) spatiotemporal masking, which is our final choice, in which masked

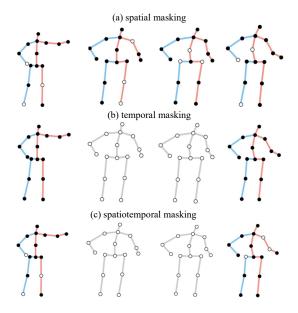


Figure 7. Illustration of three different mask sampling strategies. The joints in white/gray color represent a masked one and the joints in black color represent the unmasked one.

Mask	Hun	nan3.6M	MPI-INF-3DHP			
IVIASK	MPJPE (↓)	PA-MPJPE (↓)	PCK (↑)	AUC (↑)	MPJPE (↓)	
Spatial	47.3	37.5	51.9	84.9	84.5	
Temporal	47.7	38.0	52.2	85.2	84.3	
Spatio-temporal	46.7	37.2	52.7	85.4	83.4	

Table 8. Performance under different three mask sampling strategies on Human3.6M and MPI-INF-3DHP.

joints are picked randomly frame by frame. We conduct comparison experiments under the same masking ratio of 72.5% in total as shown in Table 8. Spatiotemporal masking outperforms the other two mask sampling strategies, leading to better relation modeling both in spatial and temporal.

4.4.4 What is the Best Masking Ratio in MPM?

Figure 8 shows the influence of the masking ratio in grid search. We claim that the masking ratio in spatial and temporal should be considered separately. The optimal ratios are $r_s^*=5/16$ in spatial and $r_t^*=60\%$ in temporal, which leads to the masking ratio of 72.5% in total. The optimal ratio is larger than 15% of BERT [9] in NLP, similar to 75% of MAE [16] in images, smaller than 95% of Video-MAE [12] in videos and 90% of P-STMO [40] in 2D pose sequences. The adjacent poses in a sequence are usually very similar, which are more redundant than words and images. Yet, modeling 2D and 3D poses together is difficult than 2D pose alone. Therefore, the optimal ratio is smaller than that of P-STMO.

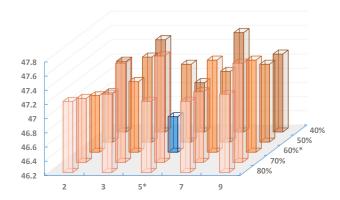


Figure 8. Grid search of masking number of joints r_s (x-axis) and ratio r_t (y-axis) for spatiotemporal masking. We report the performance over MPJPE (z-axis) on Human3.6M. The optimal ratios are $r_s^*=5/16$ in spatial and $r_t^*=60\%$ in temporal leads to 46.7 mm of MPJPE. Detailed results are listed in Table in Appendix.

5. Limitation

Due to the computational resource limitation, we could not scale up our model into billion-size datasets. Although we use multiple datasets for training and evaluation, domain shift might happen due to the poor diversity of training data.

6. Conclusion

In conclusion, this paper presents MPM, a novel framework for unifying 2D and 3D human pose representations in a shared feature space. While previous research has extensively explored estimating 3D human pose solely from 2D pose sequences, this work takes a significant step forward by integrating both modalities and leveraging a singlestream transformer-based architecture. We propose three pretext tasks: masked 2D pose modeling, masked 3D pose modeling, and masked 2D pose lifting. These tasks serve as pre-training objectives for their network, which is then fine-tuned using full-supervision. Notably, a high masking ratio of 72.5% is employed with spatiotemporal mask sampling strategy. The MPM framework offers several advantages, as it enables handling multiple tasks within a single unified framework. Specifically, it facilitates 3D human pose estimation, 3D pose estimation from occluded 2D pose data, and 3D pose completion. The effectiveness of MPM is demonstrated through extensive experiments and ablation studies conducted on multiple widely used human pose datasets. Remarkably, MPM achieves state-of-the-art performance on Human3.6M and MPI-INF-3DHP. Moreover, contrastive paradigms like CLIP [39] are also reasonable to help learning a unified representation of 2D and 3D human poses. Besides, introducing RGB or depth information as a third or fourth modality could also be considered in our future research.

Appendix

A. Detail Result in Mask Ratio

r_s v.s. r_t	0.4	0.5	0.6	0.7	0.8
2	47.2	46.9	47.0	47.1	47.2
3	47.5	47.4	47.2	47.6	47.3
5	46.9	47.3	46.7	47.7	47.2
7	47.6	47.2	47.5	47.2	47.2
9	47.3	47.3	47.5	47.6	47.3

Table A. Grid search of masking number of joints r_s and ratio r_t for spatiotemporal masking. We report the performance over MPJPE on Human3.6M. The optimal ratios are $r_s^* = 5/16$ in spatial and $r_t^* = 60\%$ in temporal leads to 46.7 mm of MPJPE.

B. Comparation with other methods in numbers of parameters

Method	Param. [M]	FLOPs [M]	MPJPE [mm]
Chen et al. [5]	58.1	116	32.3
Li et al. [25]	18.9	1030	30.5
Pavllo <i>et al</i> . [36]	17.0	34	37.8
Zheng et al. [54]	9.6	1358	31.3
Shan et al. [40]	6.7	1737	29.3
MPM(ours)	7.3	2001	27.2

Table B. Comparation with other methods in numbers of parameters.

References

- [1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Com*puter Vision and Pattern Recognition (CVPR), June 2014. 6
- [2] Wenhao Chai, Zhongyu Jiang, Jenq-Neng Hwang, and Gaoang Wang. Global adaptation meets local generalization: Unsupervised domain adaptation for 3d human pose estimation. arXiv preprint arXiv:2303.16456, 2023. 1
- [3] Wenhao Chai and Gaoang Wang. Deep vision multimodal learning: Methodology, benchmark, and trend. Applied Sciences, 12(13):6588, 2022.
- [4] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7035–7043, 2017. 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*, 32(1):198–209, 2021. 5, 6, 9
- [6] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of* the IEEE conference on computer vision and pattern recognition, pages 7103–7112, 2018. 1, 6
- [7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX, pages 104–120. Springer, 2020. 2
- [8] Hai Ci, Mingdong Wu, Wentao Zhu, Xiaoxuan Ma, Hao Dong, Fangwei Zhong, and Yizhou Wang. Gfpose: Learning 3d human pose prior with gradient fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4800–4810, 2023. 6
- [9] 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. 2, 8
- [10] Dylan Drover, Rohith MV, Ching-Hang Chen, Amit Agrawal, Ambrish Tyagi, and Cong Phuoc Huynh. Can 3d pose be learned from 2d projections alone? In *Proceedings* of the European Conference on Computer Vision (ECCV) Workshops, pages 0–0, 2018. 5
- [11] Hao-Shu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. In *Proceed*ings of the AAAI conference on artificial intelligence, volume 32, 2018. 5
- [12] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. Advances in neural information processing systems, 35:35946–35958, 2022. 2, 8
- [13] Mohsen Gholami, Bastian Wandt, Helge Rhodin, Rabab Ward, and Z Jane Wang. Adaptpose: Cross-dataset adap-

- tation for 3d human pose estimation by learnable motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13075–13085, 2022. 1
- [14] Kehong Gong, Jianfeng Zhang, and Jiashi Feng. Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8575–8584, 2021. 1, 5
- [15] Shenghao Hao, Peiyuan Liu, Yibing Zhan, Kaixun Jin, Zuozhu Liu, Mingli Song, Jenq-Neng Hwang, and Gaoang Wang. Divotrack: A novel dataset and baseline method for cross-view multi-object tracking in diverse open scenes. arXiv preprint arXiv:2302.07676, 2023. 1
- [16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2, 4, 8
- [17] Wenbo Hu, Changgong Zhang, Fangneng Zhan, Lei Zhang, and Tien-Tsin Wong. Conditional directed graph convolution for 3d human pose estimation. In *Proceedings of the 29th* ACM International Conference on Multimedia, pages 602– 611, 2021. 5, 6
- [18] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine* intelligence, 36(7):1325–1339, 2013. 1, 5
- [19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. 5
- [20] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9887–9895, 2019. 6
- [21] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2
- [22] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3316–3333, 2021.
- [23] 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*, pages 6173–6183, 2020. 6
- [24] 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. 4, 5, 6
- [25] 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, pages 13147–13156, 2022. 1, 2, 9
- [26] Jiahao Lin and Gim Hee Lee. Trajectory space factorization for deep video-based 3d human pose estimation. *arXiv* preprint arXiv:1908.08289, 2019. 5, 6
- [27] Junyang Lin, An Yang, Yichang Zhang, Jie Liu, Jingren Zhou, and Hongxia Yang. Interbert: Vision-and-language interaction for multi-modal pretraining. *arXiv preprint* arXiv:2003.13198, 2020. 2
- [28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014. 1
- [29] 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5064–5073, 2020. 5
- [30] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems, 32, 2019. 2
- [31] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. arXiv preprint arXiv:2002.06353, 2020. 2
- [32] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 5
- [33] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference* on computer vision, pages 2640–2649, 2017. 1, 2
- [34] 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. 1, 5, 6
- [35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32, 2019.
- [36] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7753–7762, 2019. 1, 2, 4, 5, 6, 7,

- [37] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. arXiv preprint arXiv:2208.06366, 2022. 2
- [38] Zhenting Qi, Ruike Zhu, Zheyu Fu, Wenhao Chai, and Volodymyr Kindratenko. Weakly supervised two-stage training scheme for deep video fight detection model. *arXiv* preprint arXiv:2209.11477, 2022. 1
- [39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 8
- [40] 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. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 461–478. Springer, 2022. 1, 2, 5, 6, 7, 8, 9
- [41] Wenkang Shan, Haopeng Lu, Shanshe Wang, Xinfeng Zhang, and Wen Gao. Improving robustness and accuracy via relative information encoding in 3d human pose estimation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3446–3454, 2021. 5
- [42] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 2
- [43] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490, 2019.
- [44] 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. 1, 3
- [45] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3d pose estimation from videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 764–780. Springer, 2020. 5
- [46] Jingwei Xu, Zhenbo Yu, Bingbing Ni, Jiancheng Yang, Xiaokang Yang, and Wenjun Zhang. Deep kinematics analysis for monocular 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on computer vision and Pattern recognition*, pages 899–908, 2020. 5
- [47] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2
- [48] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 1
- [49] Hao Yang, Dan Yan, Li Zhang, Yunda Sun, Dong Li, and Stephen J Maybank. Feedback graph convolutional network for skeleton-based action recognition. *IEEE Transactions on Image Processing*, 31:164–175, 2021.
- [50] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 507–523. Springer, 2020. 5
- [51] Dan Zeng, Yuhang Huang, Qian Bao, Junjie Zhang, Chi Su, and Wu Liu. Neural architecture search for joint human parsing and pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11385–11394, 2021. 1
- [52] 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. 6, 7
- [53] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8877–8886, 2023. 1, 2, 7
- [54] 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, pages 11656–11665, 2021. 1, 2, 5, 6, 7, 9
- [55] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF* conference on computer vision and pattern recognition, pages 8746–8755, 2020. 2