# APTPose: Anatomy-aware Pre-Training for 3D Human Pose Estimation

Qing-Wen Yang[1]
ss109062702@gapp.nthu.edu.tw
Kai-Wen Duan[1]
kevin77688@gapp.nthu.edu.tw
Ting-Yi Lu[1]
grace1287986@gapp.nthu.edu.tw
Kevin Lin[2]
keli@microsoft.com
Cheng-Yen Yang[3]
cycyang@uw.edu
Lijuan Wang[2]
lijuanw@microsoft.com
Jenq-Neng Hwang[3]
hwang@uw.edu
Shang-Hong Lai[1]
lai@cs.nthu.edu.tw

[1] Department of Computer Science National Tsing Hua University Hsinchu, Taiwan
[2] Microsoft Corporation, USA
[3] University of Washington, USA

**Abstract**

This paper presents a novel anatomy-aware pre-training method for accurate 3D human pose estimation, named APTPose. We propose a Hierarchical Masked Pose Modeling (HMPM) subtask that decouples the body skeleton into several distinct body components for hierarchical modeling. It surpasses the limitations of earlier joint coordinate masking techniques by better capturing the dependencies of the human skeletal structure. Unlike previous methods focusing on 2D pose reconstruction in their pre-training task, we leverage a large number of 3D pseudo labels from existing datasets for pre-training. This allows us to better model the skeletal system in 3D space and improve the accuracy and robustness of 3D human pose estimation. Additionally, we introduce a geometric loss into the optimization process to boost correlations within the human skeleton. Experimental results show its superior robustness and generalization capabilities across challenging benchmarks, offering a favorable balance between accuracy and computational complexity, thus making it an appealing option for practical applications. Code is available at *https://github.com/wenwen12321/APTPose*.

## 1 Introduction

Monocular 3D human pose estimation (3DHPE) is a fundamental task that involves estimating 3D poses and reconstructing body representation, such as the skeleton position,

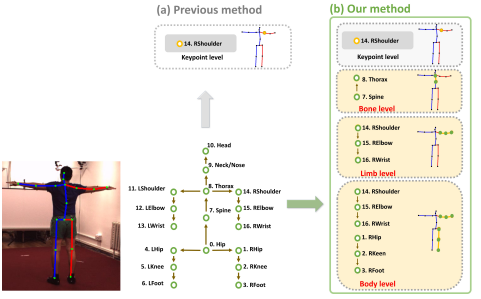



Figure 1: Illustration of the main difference between the masking strategies in APTPose and previous approach [20]. In addition to the *Keypoint-level* masking strategy, our proposed APTPose also aggregates the *Bone, Limb,* and *Body level* information into pre-training. Detailed definitions are provided in Sec. 3.1.

from a single camera. This task has a wide range of applications, including action recognition, augmented/virtual reality, and human-robot interaction, where it is instrumental in capturing complex body geometry and motion expressions. Moreover, the current trend in 3DHPE is revolved around the 2D-to-3D lifting pipeline, which uses 2D keypoints from off-the-shelf detectors [2, 7] as input and focusing on lifting 3D pose. To enhance accuracy, many sequence-based methods [1, 18, 31, 33] leverage temporal information from videos to achieve notable improvements over single-frame methods.

Later, with the superior performance of the Transformer [25], recent works [15, 16, 31, 32, 33] have incorporated the Transformer architecture for further advancements in pose fields. Shan *et al*. [20] proposed a pre-training model for 2D-to-3D human pose estimation, leading to significant advancements in performance. The model effectively captures spatial and temporal dependencies by randomly masking both spatial and temporal domains from input 2D sequences and recovering corrupted 2D poses using a denoising auto-encoder. However, we note that there are still several limitations in their approach: (*i*) Masking strategy employed in previous approach only considers Keypoint-level masking in the spatial domain, overlooking the rich human skeleton structural information and oversimplifying the pose estimation task. (*ii*) Our second concern is that it predominantly concentrates on 2D pose reconstruction in pre-training, disregarding the significance of depth and motion information in 2D-to-3D lifting task. (*iii*) In addition, the method lacks sufficient geometric knowledge to enable models to comprehend the relationships between human skeleton. Consequently, these reasons highlight the limitations of the previous pre-training method.

Driven by these observations and analysis, we present a novel **A**natomy-aware **P**re-**T**raining approach, dubbed APTPose, that leverages anatomical knowledge for human pose estimation. To address the first issue, we introduce a Hierarchical Masked Pose Modeling (HMPM), which decouples the body skeleton into several distinct body components, including *Keypoint-level, Bone-level, Limb-level* and *Body-level* as shown in Figure 1 for effective hierarchical pose representation learning. To address the second issue, APTPose effectively combines 2D and 3D supervision by a simple auxiliary reprojection module. This module seamlessly integrates an extensive set of 3D pseudo-labels derived from widely available in-the-wild 2D data, enabling the precision and robustness of the human skeletal structure representation. Furthermore, APTPose tackles the third issue by introducing geometric loss constraint into the optimization process. This loss function inherently carries the essential geometric knowledge, enabling our model to estimate plausible poses by considering the

bone orientation and bone length characteristics of the human pose.
In summary, we make the following contributions:

- We present an anatomy-aware pre-training framework to discern human skeletal structure across distinct body components effectively.
- We utilize a reprojection module to combine 2D and 3D supervision in pre-training, enabling the precision and robustness of the human skeletal structure representation.
- We introduce a geometric loss into the optimization process to further improve the consistency and plausibility of pose predictions.

## 2 Related Works

In the interest of space, we limit our discussion to prior work on video-based 2D-to-3D lifting approaches in single-person and single-view settings. Some of the early studies [1, 4, 18, 34] leveraged temporal information from the adjacent frames to mitigate depth ambiguity. [18] proposed a temporal convolutional network (TCN) that integrates spatial-temporal dependencies across sequences. [1] decomposed the estimation task into bone length and bone direction prediction subtasks, effectively capturing information from both local and distant frames. However, these works rely on simple operations to map local joints coordinate to a latent space, neglecting the long-range dependencies with temporal connectivity. To address this problem, [33] employed the Transformer architecture to more effectively model spatial-temporal information.[15] further explored incorporating stride 1D convolution into the Transformer architecture to aggregate the full sequence into a central frame.

Although existing 2D-to-3D models achieve impressive performance under controlled laboratory settings, their effectiveness on real-world is frequently hindered by the scarcity and lack of diversity in publicly available training data. Researchers [3, 8, 14] have tried using data augmentation strategy to generate diverse 2D-3D pose pairs. [14] adopted evolutionary operators to generate variations of 3D poses. [8] designed an end-to-end generative model to create plausible new poses. However, these methods are still limited to the combinations derived from *"seen"* indoor source data, making it challenging to infer accurate results in real-world *"unseen"* scenarios.

To further enhance generalization to in-the-wild scenarios, some studies [12, 26, 27, 29] have explored self-supervised learning methods to train models with unlabeled data. Meanwhile, extensive works [9, 23, 29, 34] have incorporated mixed 2D in-the-wild data in the training process using weakly-supervised learning or transfer learning techniques. Inspired by these studies, we explore the utilization of existing in-the-wild 2D annotations datasets to align with real-world scenarios, leading to marginal boost to the model generalization.

In recent years, transformer-based pre-training methods have emerged as crucial solutions in various fields to cope with limited data availability, significantly enhancing model robustness and generalization. In natural language processing, [6, 19, 24] proposed masked language modeling (MLM), randomly masking words and predicting masked words based on context. Shifting to computer vision, [10] delves into masked image modeling (MIM), involving the random masking of pixels or visual tokens and subsequent reconstruction of missing regions. Moreover, the influence of pre-training techniques has expanded to areas like video processing [22, 28] and video-language tasks [21]. Along this line, [20] pioneered the application of pre-training techniques to 3DHPE, introducing masked pose modeling (MPM). The task involves inpainting masked joints in the spatial domain and filling in masked frames in the temporal domain. In contrast to existing MPM [20] task, our proposed hierarchical masked pose modeling (HMPM) considers deeper dependencies within

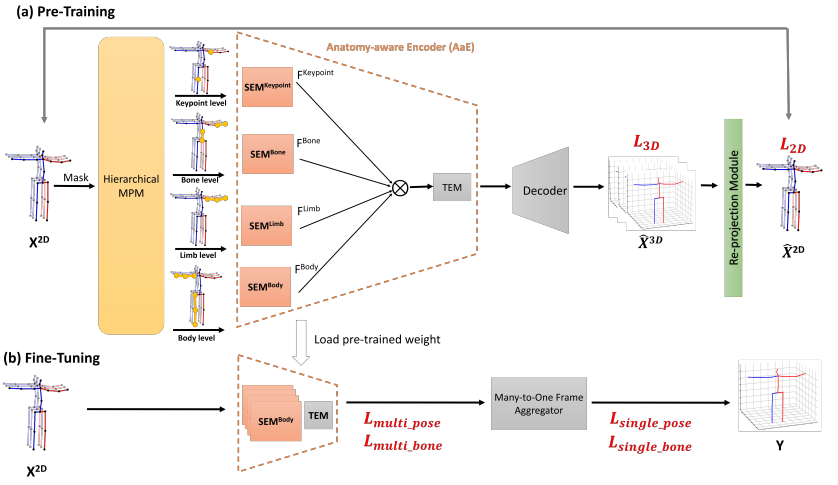



Figure 2: **Overview of our proposed APTPose. (a) Pre-training stage:** Hierarchical MPM masks the given 2D pose sequence at different anatomy levels, which is then fed to the Encoder. Decoder reconstructs 3D pose sequence from encoded unmasked embeddings and temporal padding embedding. Reprojection module projects 3D pose sequence onto 2D plane. **(b) Fine-tuning stage.** The 2D pose sequence is fed to the pre-trained Encoder and Aggregator for estimating full sequence and single frame of 3D pose, respectively.

the human skeletal structure, leading to improved accuracy and generalization of prediction.

# 3 Proposed Method

Figure 2 depicts an overview of the APTPose. Our proposed method has a two-stage pipeline, consisting of pre-training and fine-tuning. The goal of the pre-training phase is to enable our encoder to capture human pose representation across spatial-temporal relationships with the help of HMPM. Subsequently, fine-tuning is dedicated to lifting 2D poses to their corresponding 3D pose. This approach facilitates optimization and achieves strong performance.

During the pre-training stage, our model tends to effectively learn hierarchical pose representations at each anatomy level by solving the proposed **H**ierarchical **M**asked **P**ose **M**odeling (HMPM) subtask. We utilize an **A**natomy-**a**ware **E**ncoder (AaE) and Decoder to reconstruct both 2D and 3D pose sequences. The AaE consists of four Spatial Encoding Modules (SEMs) and a Temporal Encoding Module (TEM), designed to capture the specific level of anatomy-aware human skeletal structure and long-range temporal dependencies, respectively. For computational efficiency, SEM is constructed using an MLP block instead of Transformer architecture as in TEM. Additionally, we enhance the learning process by integrating both 2D and 3D supervision through a straightforward auxiliary reprojection module. It significantly improves the precision and robustness of pose estimation by enabling the model to learn spatial and temporal relationships across both 2D and 3D domains, as well as their complex mapping relationships. More details of the model architecture can be found in the supplementary. After pre-training, we fine-tune the pre-trained AaE and Aggregator to estimate a 3D pose. In addition to the typical optimization process, we introduce geometric loss constraints, enhancing the consistency and plausibility of pose estimations.

### 3.1 Hierarchical MPM

**Definition.** We decouple the human skeleton into several distinct body components for hierarchical modeling. These levels start from local to global representations, including (1) *Keypoint level*, (2) *Bone level*, (3) *Limb level*, and (4) *Body level* as illustrated in Figure 1. Firstly, we denote *Keypoint level* as each individual joint $J$ as defined in previous work [20]. We define *Bone level* as any two neighboring joints where they essentially formed $J-1$ bone (*e.g.*, *Thorax-Spine*). *Limb level* representes the four limbs of the human body, which are represented by three consecutive joints on arms and legs (*e.g.*, *Shoulder-Elbow-Wrist* and *Hip-Knee-Foot*). Lastly, *Body level* represents either the right side or the left side of the body (*e.g.*, *RShoulder-RElbow-RWrist, RHip-RKnee-RFoot*). These four masking variations aim to exploit the relations of sets of keypoints, prompting the encoder to effectively learn from representation of human anatomy.

**Process of Hierarchical MPM.** The implementation details of our HMPM are shown in Figure 2 (a). We adopt the HMPM approach by randomly masking a fixed set of 2D joints in each frame and filling the masked joints with learnable shared parameters for each masking level strategy in parallel. Specifically, we require each masking level to output a corresponding spatial encoder input, *i.e.*, $X^{Keypoint}, X^{Bone}, X^{Limb}, X^{Body} \in \mathbb{R}^{N\times 2J}$. As these masking strategies share the same implementation, we present the *Bone level* masking as an example to illustrate our approach. The procedure can be formulated as Equation 1:

$$\begin{aligned} X^{Bone} &= \{x_1^{Bone}, x_2^{Bone}, \dots, x_N^{Bone}\}, \\ x_n^{Bone} &= \{p_i^{2D} : i \notin M_n^{Bone}\}_{i=1}^{J} \bigcup \{e^{Bone} : i \in M_n^{Bone}\}_{i=1}^{J} \end{aligned} \tag{1}$$

where $M_n^{Bone} \in \mathbb{R}^{N\times |m^{Bone}|}$ represents the masked joints at the bone level in frame $n$, and $m^{Bone} \in \mathbb{N}$ denotes the number of masked joints at the bone level, $e^{Bone} \in \mathbb{R}^{m^{Bone}}$ contains shared learnable parameters for padding the masked joints at the bone level, and $x_n^{Bone}$ is a 2D pose with bone masking. Note that $x_n^{Bone}$ is formed by replacing the 2D pose $p_i^{2D}$ for the joints with bone masking by padding.

As a result, each SEM learns distinct body components representation by feeding $X^{Keypoint}$, $X^{Bone}, X^{Limb}, X^{Body}$ as input and outputting $F^{Keypoint}, F^{Bone}, F^{Limb}, F^{Body} \in \mathbb{R}^{J\times d}$, respectively, where $d$ denotes the dimension of latent representation.

Subsequently, we obtain the hierarchical spatial features $F^H \in \mathbb{R}^{J\cdot d}$ by a simple yet effective hierarchical feature fusion module, which performs element-wise multiplication to fuse the representations learned from different levels.

### 3.2 Reprojection Module and Noising Pipeline

In image-based pose estimation, 2D/3D mixed data [23, 29] and 3D pseudo annotations [5, 13] are commonly used to enhance data diversity. For image inputs, performance gains primarily result from a diversity of appearances, whereas for 2D-to-3D lifting, they arise from incorporating various mapping relationships between 2D and 3D poses. In this section, our key insight is to achieve two objectives using a simple Reprojection Module and Noising Pipeline: (i) extending the optimization process from 2D-only to both 2D and 3D, (ii) enabling the integration of easily accessible image-based 2D in-the-wild datasets into our frame-input framework. Details of these performance improvements are reported in Table 2.

**Reprojection Module.** To combine 2D and 3D supervision during pre-training, our proposed reprojection module projects the predicted 3D poses from the decoder into 2D space,

using a combination of MLP and residual networks. Moreover, this module also allows our framework to seamlessly integrate an extensive set of 3D pseudo-labels derived from widely available in-the-wild 2D data. Experimental results demonstrate that integrating 3D supervision during pre-training positively impacts the learned representations.

**Noising Pipeline.** We also conduct experiments on incorporating existing in-the-wild 2D dataset (COCO) and 3D dataset (H36M) for pre-training. To address the discrepancy between the image-based format and our sequence-based framework, we introduce a nosing pipeline. This involves duplicating images into the required frames and adding random noise and jitter to pose annotations for sequence simulation, thereby enhancing generalization by augmenting training scenarios. It is worth noting that we use the 3D pseudo labels of 2D dataset provided by [5]. For further details on the noising pipeline, please refer to the supplementary material.

## 3.3 Geometric Knowledge Constraints

Rather than arbitrarily predicting coordinates, we aim to incorporate the geometric constraints of the correlations in the human skeleton to improve our training process. To achieve this purpose, we explore previous geometric constraint methods [8, 14, 26] and note that bone vectors can potentially convey more comprehensive geometric information (*e.g.*, *bone orientation* and *bone length*) than individual joint locations. We define *Pelvis* as the root joint of 3D coordinate, and represent our 3D pose as the root-relative human skeleton consisting of joints (as nodes) and bone lengths (as edges), shown in the Figure 1, where each yellow arrow pointing from parent joints to child joints can be viewed as a single bone vector. Note that each joint $j$ has only one corresponding parent joint, denoted by $parent(j)$. Given a set of 3D joint locations that contains $J$ joints in $i$-th frame $x_i^{3D} = \{p_1^{3D}, ..., p_J^{3D}\}$, $x_i^{3D} \in \mathbb{R}^{(J)\cdot 3}$, we can acquire $(J-1)$ corresponding bone vectors $B_i = \{b_1, ..., b_{J-1}\}$, $B_i \in \mathbb{R}^{(J-1)\cdot 3}$ by substracting the parent joint corresponding to joint $j$. For example, the $j^{th}$ bone vector $b_j$ can be defined as follows:

$$b_j = p_j^{3D} - p_{parent(j)}^{3D}. \tag{2}$$

## 3.4 Loss Functions

Our proposed model is trained by using a two-stage pipeline, consisting of pre-training and fine-tuning. We detail the loss functions used in each stage below.

**Stage I. Pre-training:** The pre-training objective of our model, denoted as $L_{pre-train}$, comprises two supervised loss components: the 3D pose loss and the 2D reprojection loss. We employ standard L2 loss to minimize the difference between the predicted and ground truth poses:

$$L_{pre-train} = \lambda_{3D} L_{3D} + \lambda_{2D} L_{2D}, \tag{3}$$

where $L_{3D}$ and $L_{2D}$ are 3D pose and 2D pose losses, respectively. We use weighting factors $\lambda_{3D}$ and $\lambda_{2D}$ to balance the effect of these two loss functions.

**Stage II. Fine-tuning:** Similar to [15, 20], our fine-tuning stage adopts a full-to-single prediction scheme. The multi-frame loss first guides the TEM in leveraging temporal relationships across the full sequence to enforce temporal smoothness. Subsequently, the single-frame loss refines the representation by aggregating these 3D features from the sequence to estimate the more accurate 3D pose of the center frame. As mentioned in Sec. 3.3, we incorporate bone vector into the loss function along with the loss for both multi-frame and single-frame training. Therefore, there are four types of supervision in this stage. For all of

these cases, $L_2$ norm is employed as the loss function for the prediction error with respect to the ground truth. The multi-frame pose and bone vector supervision losses, denoted by $L_{multi\_pose}$ and $L_{multi\_bone}$, respectively, are defined as:

$$L_{multi\_pose} = \sum_{n=1}^{N}\sum_{j=1}^{J}\left\|p_{j,n}^{3D} - \hat{p}_{j,n}^{3D}\right\|_2 \;\;\&\;\; L_{multi\_bone} = \sum_{n=1}^{N}\sum_{j=1}^{J-1}\left\|b_{j,n} - \hat{b}_{j,n}\right\|_2, \tag{4}$$

where $p_{j,n}^{3D}$ and $\hat{p}_{j,n}^{3D}$ denote the $j$-th predicted and ground truth 3D joint locations in the $n$-th frame, respectively, $b_{j,n}$ and $\hat{b}_{j,n}$ denote the $j$-th predicted and ground truth 3D bone vectors in the $n$-th frame, respectively, $N$ denotes the total number of frames, and $J$ denotes the total number of joints. The single-frame pose and bone vector supervision losses, denoted by $L_{single\_pose}$ and $L_{single\_bone}$, respectively, are defined as:

$$L_{single\_pose} = \sum_{j=1}^{J}\left\|p_{j}^{3D} - \hat{p}_{j}^{3D}\right\|_2 \;\;\&\;\; L_{single\_bone} = \sum_{j=1}^{J-1}\left\|b_{j} - \hat{b}_{j}\right\|_2, \tag{5}$$

where $p_j^{3D}$ *and* $\hat{p}_j^{3D}$ denote the $j$-th predicted and ground truth 3D joint locations, respectively, and $b_j$ and $\hat{b}_j$ denote the $j$-th predicted and ground truth 3D bone vectors, respectively. Overall the total loss function $L_{total}$ for model fine-tuning can be written as follows:

$$L_{total} = \lambda_{m_p} L_{multi\_pose} + \lambda_{m_b} L_{multi\_bone} + \lambda_{s_p} L_{single\_pose} + \lambda_{s_b} L_{single\_bone} \tag{6}$$

where $\lambda$'s are the weighting factors to balance the above loss functions.

# 4 Experimental Results

## 4.1 Datasets and Evaluation Metrics

**Human 3.6M (H36M)** [11] is the largest indoor 3DHPE benchmark with 3.6 million video frames covering 15 activities by 11 subjects. We train on subjects S1, S5, S6, S7, S8, and evaluate on S9 and S11. Evaluation metrics include Mean Per-Joint Position Error (MPJPE) and Procrustes analysis MPJPE (P-MPJPE) in millimeters.

**MPI-INF-3DHP (3DHP)** [17] is another substantial benchmark for 3DHPE, encompassing both indoor and outdoor scenes with 1.3 million frames. It presents more diverse and challenging motions than H36M. The Percentage of Correct Keypoints (PCK), Area Under the Curve (AUC), and MPJPE are reported as evaluation metrics.

## 4.2 Comparison with State-of-the-Art Methods

We compare our proposed APTPose against state-of-the-art methods using the H36M and 3DHP datasets, with results detailed in Table 1a and Table 1b, respectively. Employing a similar model architecture design, our method surpasses the previous pre-trained baseline, P-STMO [20], by incorporating the enhancements outlined in our introduction. On H36M, our method achieves improvements of 0.46% in MPJPE and 1.16% in P-MPJPE. On 3DHP, our method outperforms [20] by 4.3% in MPJPE, 1.3% in AUC, and 0.1% in PCK. In addition, A key observation is APTPose's exceptional performance in handling shorter sequence lengths (with frames $f \leq 27$) as illustrated in Table 1b. This advantage is mainly due to the integration of geometric constraints in our model, which ensures plausible pose estimation even with limited observational data. Such an approach is effective in scenarios typical of online and daily life videos, which are often characterized by their fast-paced

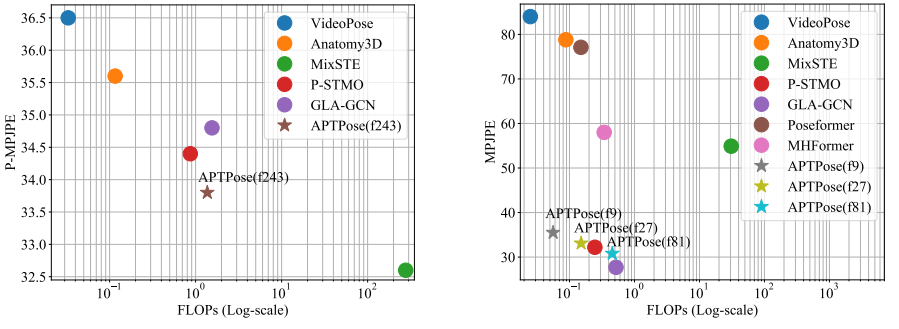
| Method | $f$ | FLOPs(G)↓ | MPJPE↓ | P-MPJPE↓ |
|---|---|---|---|---|
| VideoPose [CVPR'19] | 243 | 0.033 | 46.8 | 36.5 |
| Anatomy3D [TCSVT'21] | 243 | 0.116 | 44.1 | 35.6 |
| MixSTE [CVPR'22] | 243 | 277.24 | 40.9 | 32.6 |
| P-STMO [ECCV'22] | 243 | 0.868 | 42.8 | 34.4 |
| GLA-GCN [ICCV'23] | 243 | 1.558 | 44.4 | 34.8 |
| **APTPose** | 243 | 1.367 | 42.6 | 34.0 |
| **APTPose (+Extra2D)** | 243 | 1.367 | 42.5 | 33.8 |

(a) Results on **H36M**.

| Method | $f$ | FLOPs(G)↓ | PCK↑ | AUC↑ | MPJPE↓ |
|---|---|---|---|---|---|
| Poseformer [ICCV'21] | 9 | 0.15 | 88.6 | 56.4 | 77.1 |
| MHFormer [CVPR'22] | 9 | 0.342 | 93.8 | 63.3 | 58.0 |
| **APTPose** | 9 | 0.056 | 97.1 | 73.8 | 35.5 |
| MixSTE [CVPR'22] | 27 | 30.8 | 94.4 | 66.5 | 54.9 |
| **APTPose** | 27 | 0.151 | 97.4 | 75.2 | 33.1 |
| VideoPose [CVPR'19] | 81 | 0.025 | 86.0 | 51.9 | 84.0 |
| Anatomy3D [TCSVT'21] | 81 | 0.088 | 87.9 | 54.0 | 78.8 |
| P-STMO [ECCV'22] | 81 | 0.246 | 97.9 | 75.8 | 32.2 |
| GLA-GCN [ICCV'23] | 81 | 0.520 | 98.5 | 79.1 | 27.7 |
| **APTPose** | 81 | 0.455 | 98.0 | 76.8 | 30.8 |
| **APTPose (+Extra2D)** | 81 | 0.455 | 97.7 | 77.2 | 30.5 |

(b) Results on **3DHP**.

Table 1: **(a) Results on H36M dataset.** 2D poses detected by CPN are used as inputs. **(b) Results on 3DHP dataset.** 2D GT keypoints are used as inputs. The best results are highlighted in red. The second-best are highlighted in blue.

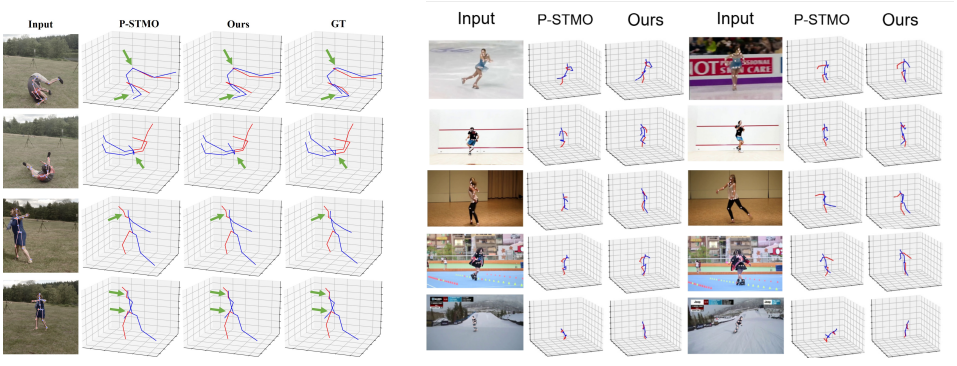



(a) H36M dataset (b) 3DHP dataset

Figure 3: **Comparison of computational cost** between APTPose and other SOTA methods on two benchmark datasets.

and brief nature. However, while MixSTE demonstrates the lowest reconstruction error on H36M, it incurs a significant computational overhead that exceeds other methods by over *200 times*. Moreover, its performance in the in-the-wild scenarios of 3DHP is suboptimal, leading us to characterize it as overfitted. On the other hand, GLA-GCN exhibits strong generalization capabilities on challenging 3DHP datasets when supplied with high-quality 2D data (*i.e.*, ground truth poses). Nonetheless, its performance deteriorates considerably when low-quality 2D data is used (*i.e.*, 2D poses detected by the CPN detector), even on relatively simpler datasets like H36M. In contrast, APTPose achieves competitive accuracy with significantly reduced computational demands. This efficiency is attributed to our proposed pre-training strategy, which enables the lightweight Anatomy-aware Encoder (AaE) to effectively capture human skeletal structures. Additionally, our masking strategy during pre-training enhances robustness against noise, making APTPose less sensitive to variations in 2D pose quality. Building on these observations, APTPose demonstrates competitive performance across various scenarios by effectively balancing complexity, noise robustness, stability, and generalization capabilities. These characteristics highlight its suitability for real-world applications.

**Qualitative Results.** As shown in Figure 4a, we present a qualitative comparison between our proposed method and [20] on 3DHP dataset. By leveraging anatomy-aware knowledge, our method yields greater consistency between the estimated bone lengths and the GT. This improvement contributes to more accurate pose estimation, particularly in complex outdoor




(a) 3DHP dataset (b) In-the-wild videos

Figure 4: Qualitative Results on challenging outdoor scenarios.

scenes compared to [20]. We will discuss this further in the upcoming ablation study. To evaluate the generalization capabilities of our model, we employ the model weights trained on the H36M dataset for conducting in-the-wild video inference. As illustrated in Figure 4b, our proposed hierarchical masking strategy demonstrates superior performance in real-world scenarios compared to [20]. The results distinctly reveal that APTPose can better model the human skeletal structure, facilitating enhanced generalization to challenging and uncontrolled environments. Please refer to the supplementary material for additional quantitative and qualitative evaluation results.

## 4.3 Ablation Studies

**Effectiveness of Individual Components.** In Table 2a, we progressively integrate our proposed components into the primitive backbone on the H36M dataset. Initially, we incorporate geometric loss to enhance the model's ability of capturing correlations within the human skeleton, leading to a 5.6% improvement in accuracy compared to the baseline. Subsequently, we evaluate the impact of HMPM strategy under two pre-training supervision settings: (*i*) 2D only, and (*ii*) combined 2D and 3D supervision. The addition of HMPM during pre-training reduces MPJPE from 43.5 mm to 42.9 mm. Furthermore, the integration of 3D supervision during pre-training yields even more pronounced benefits, reducing MPJPE to 42.6 mm. The results emphasize the critical role of 3D supervision in pre-training, highlighting the limitations of [20] focused on 2D representations.

**Effectiveness of Geometric Constraints.** We assess the impact of our geometric constraints by examining bone length errors between predicted and ground truth skeletons on the 3DHP test set (TS1 to TS6). Table 2b demonstrates the efficacy of our geometric loss, leading to reduced bone length errors across TS2, 3, 4, and 5, thus yielding a lower average error compared to prior works [20].

**Effectiveness of Pre-Training.** In Table 2c and Table 2d, we examine the performance gains achieved by transitioning from a fundamental (keypoint-level) masking strategy to a hierarchical masking strategy on the H36M dataset, utilizing CPN as the 2D detector. To ensure a fair comparison with prior work [20], which employs only 2D loss during the pre-training stage, our results also focus on the 2D loss (MPJPE) during pre-training. Additionally, to simplify model complexity, all experiments are conducted using a 9-frame setting. Table 2c presents an ablation study that evaluates the effectiveness of different spatial masking levels in HMPM. Starting with a keypoint-level masking approach (as used in [20]), the method

| Backbone | Geo. loss | HMPM | 3D loss | MPJPE |
|---|---|---|---|---|
| ✓ | | | | 46.1 |
| ✓ | ✓ | | | 43.5 |
| ✓ | ✓ | ✓ | | 42.9 |
| ✓ | ✓ | ✓ | ✓ | **42.6** |

(a) Effectiveness of each component

| | TS1 | TS2 | TS3 | TS4 | TS5 | TS6 | Avg. |
|---|---|---|---|---|---|---|---|
| P-STMO[41] | 6.68 | 4.90 | 4.56 | 5.29 | 2.56 | 1.75 | 4.29 |
| Ours | 6.74 | **4.73** | **4.42** | **5.17** | **2.13** | 1.77 | **4.16** |

(b) Bone length error

| Keypoint | Bone | Limb | Body | MPJPE |
|---|---|---|---|---|
| ✓ | | | | 22.46 |
| | ✓ | | | 22.29 |
| | | ✓ | | 21.30 |
| | | | ✓ | 21.71 |
| ✓ | ✓ | ✓ | ✓ | **20.28 (Ours)** |

(c) Different Masking Levels

| Fusion | MPJPE |
|---|---|
| Add | 23.23 |
| Avg. | 23.19 |
| Cat. | 22.81 |
| **Mul.** | **20.28** |

(d) Different Fusions

| FT on 3DHP | MPJPE |
|---|---|
| w/o PT | 32.0 |
| PT on H36M (2D loss) | 31.3 |
| PT on H36M+COCO (2D loss) | 31.6 |
| PT on H36M+COCO (2D+3D loss) | **30.5** |

(e) Pre-training with Different Data

Table 2: **(a)** Ablation study on the effectiveness of each component (measured by MPJPE). **(b)** Ablation study on the effectiveness of geometric constraints. The unit of the bone length error is milimeter (mm). **(c)-(d)** Ablation studies on the effectiveness of our HMPM pre-training strategy, focusing on the 2D loss (MPJPE) during the pre-training stage. **(e)** Ablation study on the effectiveness of incorporating 3D supervision and additional 2D data in the pre-training stage. (**PT** denotes pre-training, **FT** denotes fine-tuning).

achieves an MPJPE of 22.46 mm. Incremental improvements in reconstruction error are observed with the implementation of bone-level, limb-level, and body-level masking strategies, resulting in MPJPE reductions of 0.17 mm, 1.16 mm, and 0.75 mm, respectively. When all masking levels are integrated, the MPJPE is further reduced to 20.28 mm. Table 2d shows a comparison of different hierarchical feature fusion mechanisms employed in our HMPM pre-training approach. Our results demonstrate that element-wise multiplication effectively incorporates human skeleton prior knowledge of different levels, whereas other mechanisms such as addition, averaging, and concatenation do not achieve comparable improvements. Table 2e presents a comprehensive analysis of the effect of incorporating 3D supervision and additional 2D data in the pre-training stage. We consider four scenarios: (*i*) no pre-training, (*ii*) pre-trained on H36M (using 2D loss only), (*iii*) pre-trained on both H36M and COCO (using 2D loss only), and (*iv*) pre-trained on both H36M and COCO (using both 2D and 3D losses). Notably, we do not include the 3DHP dataset in the pre-training stage, but only fine-tune our model on it. The last two rows in Table 2e demonstrate that our proposed strategy for handling image data and pseudo-labels allows existing 2D image datasets to be compatible with our sequence model and leads to a significant improvement.

## 5 Conclusion

We present a novel anatomy-aware pre-training framework to model comprehensive representation of the human skeletal structure by leveraging the distinct body component features. Moreover, by integrating 3D supervision and geometric knowledge constraints into the optimization process, our model significantly improves accuracy and enables more plausible 3D pose estimation. Extensive experimental results show that the proposed method has a fundamental advantage over *Keypoint-level* pre-training model, not only demonstrating its superior robustness and generalization capabilities to challenging benchmark and also providing a favorable trade-off between accuracy and computational complexity, making it a compelling choice for practical applications.

# References


[1] Tianlang Chen, Chen Fang, Xiaohui Shen, Yiheng Zhu, Zhili Chen, and Jiebo Luo. Anatomy-aware 3d human pose estimation with bone-based pose decomposition. *arXiv preprint arXiv:2002.10322*, 2020.

[2] 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.

[3] Ziyi Chen, Akihiro Sugimoto, and Shang-Hong Lai. Learning monocular 3d human pose estimation with skeletal interpolation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4218–4222. IEEE, 2022.

[4] Yu Cheng, Bo Yang, Bo Wang, Wending Yan, and Robby T Tan. Occlusion-aware networks for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 723–732, 2019.

[5] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 769–787. Springer, 2020.

[6] 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.

[7] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2334–2343, 2017.

[8] 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.

[9] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Gerard Pons-Moll, and Christian Theobalt. In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10905–10914, 2019.

[10] 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.

[11] 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.

[12] Umar Iqbal, Pavlo Molchanov, and Jan Kautz. Weakly-supervised 3d human pose learning via multi-view images in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[13] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2252–2261, 2019.

[14] 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 *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *Technical report*, 2018.

[20] 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.

[21] Bin Shao, Jianzhuang Liu, Renjing Pei, Songcen Xu, Peng Dai, Juwei Lu, Weimian Li, and Youliang Yan. Hivlp: Hierarchical interactive video-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13756–13766, 2023.

[22] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019.

[23] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European conference on computer vision (ECCV)*, pages 529–545, 2018.

[24] Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*, 2021.

[25] 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.

[26] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7782–7791, 2019.

[27] Bastian Wandt, Marco Rudolph, Petrissa Zell, Helge Rhodin, and Bodo Rosenhahn. Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[28] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14733–14743, 2022.

[29] Cheng-Yen Yang, Jiajia Luo, Lu Xia, Yuyin Sun, Nan Qiao, Ke Zhang, Zhongyu Jiang, Jenq-Neng Hwang, and Cheng-Hao Kuo. Camerapose: Weakly-supervised monocular 3d human pose estimation by leveraging in-the-wild 2d annotations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2924–2933, January 2023.

[30] Bruce XB Yu, Zhi Zhang, Yongxu Liu, Sheng-hua Zhong, Yan Liu, and Chang Wen Chen. Gla-gcn: Global-local adaptive graph convolutional network for 3d human pose estimation from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8818–8829, 2023.

[31] 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.

[32] Weixi Zhao, Weiqiang Wang, and Yunjie Tian. Graformer: Graph-oriented transformer for 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20438–20447, June 2022.

[33] 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.

[34] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *Proceedings of the IEEE international conference on computer vision*, pages 398–407, 2017.