arXiv:2404.09499v1 [cs.CV] 15 Apr 2024

# Learning Human Motion from Monocular Videos via Cross-Modal Manifold Alignment

**Shuaiying Hou**
State Key Lab of CAD&CG
Zhejiang University

**Hongyu Tao**
State Key Lab of CAD&CG
Zhejiang University

**Junheng Fang**
Bournemouth University

**Changqing Zou**
State Key Lab of CAD&CG
Zhejiang University

**Weiwei Xu***
State Key Lab of CAD&CG
Zhejiang University

## Abstract

Learning 3D human motion from 2D inputs is a fundamental task in the realms of computer vision and computer graphics. Many previous methods grapple with this inherently ambiguous task by introducing motion priors into the learning process. However, these approaches face difficulties in defining the complete configurations of such priors or training a robust model. In this paper, we present the Video-to-Motion Generator (VTM), which leverages motion priors through cross-modal latent feature space alignment between 3D human motion and 2D inputs, namely videos and 2D keypoints. To reduce the complexity of modeling motion priors, we model the motion data separately for the upper and lower body parts. Additionally, we align the motion data with a scale-invariant virtual skeleton to mitigate the interference of human skeleton variations to the motion priors. Evaluated on AIST++, the VTM showcases state-of-the-art performance in reconstructing 3D human motion from monocular videos. Notably, our VTM exhibits the capabilities for generalization to unseen view angles and in-the-wild videos.

## 1 Introduction

Modeling 3D human motion from 2D inputs, e.g. 2D keypoint sequences or monocular videos, is a fundamental cornerstone for various computer vision and computer graphics tasks, such as human action recognition, behavior analysis, video games, and virtual/augmented reality, rtc. However, this endeavor presents considerable challenges due to the inherent ambiguities in inferring 3D from 2D. Imposing motion priors has proven an effective strategy to mitigate ambiguities and increase the 3D pose plausibility.

Previous works on modeling motion priors can be broadly categorized into two groups: explicit and implicit prior based methods. Explicit methods primarily focus on restricting joint angles based on biomechanics Hatze (1997); Kodek and Munih (2002) or directly optimizing observed data Akhter and Black (2015). Unfortunately, the full configuration of pose-dependent joint angle limitations for the whole body remains unknown. Implicit methods include those modeling the plausible motion distribution with Gaussian mixture model Bogo et al. (2016); Hou et al. (2023b), variational autoencoders (VAEs) Pavlakos et al. (2019); Ling et al. (2020), generative adversarial networks (GANs) Peng et al. (2021); Davydov et al. (2022) or denoising score matching (DSM) Ci et al. (2023). These models either suffer from issues like posterior/mode collapse or difficulty in training.

*Corresponding Author

Preprint. Under review.

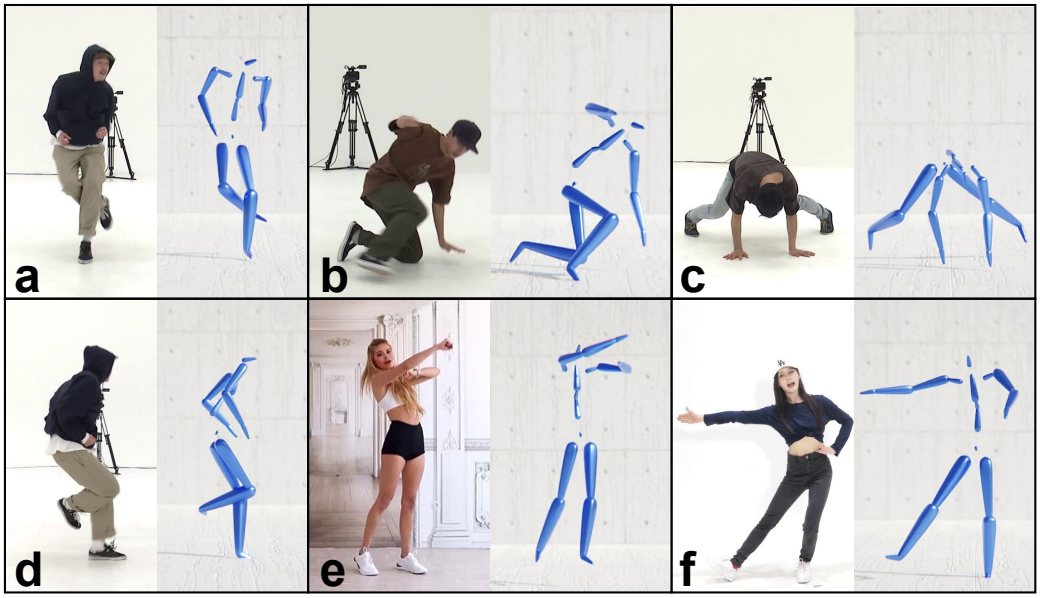


Figure 1: Our VTM can reconstruct 3D human motion from a wide range of monocular videos. These poses can be reconstructed from: a), b) and c), three frames from the validation dataset, d) the frame containing the same pose as that in a) but with an unseen view-angle during training, e) and f), two frames from the internet videos.

In light of the success of the contrastive language-image pre-training (CLIP) model Radford et al. (2021), we take a new perspective in harnessing 3D motion priors to aid the reconstruction of human motion from monocular videos by aligning the latent feature space of 3D skeleton motion data and video data. As several datasets, such as Li et al. (2021b); Ionescu et al. (2013), contain the 2D videos of human performance and their corresponding 3D motion, it's possible to find a compact and well-defined latent manifold shared by data from these two domains, similar to the CLIP latent manifold. This latent space should contain ample information to reconstruct human poses faithfully. Therefore, there are two technical challenges when performing cross-modal alignment in 3D human pose reconstruction: 1) identifying such a latent manifold, which encapsulates the kinematic constraints embodied in 3D human motion and serves as our motion priors, and 2) finding an effective way to align video feature space to the motion priors to achieve high-fidelity 3D motion reconstruction.

To address the aforementioned two challenges, we introduce a novel neural network called VTM. Firstly, the VTM employs a two-part motion auto-encoder (TPMAE), which segregates the learning of the motion latent manifold into the upper and lower body parts. This division effectively reduces the complexity of modeling the entire human pose manifold. The TPMAE is trained on 3D human motion data aligned with a scale-invariant virtual skeleton, which is beneficial for eliminating the impact of the skeleton scale on the manifold. Subsequently, the VTM incorporates a two-part visual encoder (TPVE) to translate visual features, composed of video frames and 2D keypoints, into two streams of visual latent manifolds for the upper and lower body parts. A manifold alignment loss is employed to pull the cross-modal manifolds for the upper and lower body parts closer. Finally, the TPVE is jointly trained with the pre-trained TPMAE to reconstruct 3D human motion with a complete representation: the rotations for all joints, the translations for the root joint, and a kinematic skeleton containing the scale information. This meticulously designed process ensures harmonization between the motion and visual representations within the VTM framework.

We comprehensively evaluate our VTM framework on the AIST++ dataset Li et al. (2021b). The results demonstrate that our VTM surpasses or performs comparably to state-of-the-art (SOTA) methods in terms of mean per joint position error (MPJPE). Notably, VTM exhibits the capability to swiftly generate precise human motion in sync with video sequences (~70fps). This real-time performance positions it as a potential candidate for various real-time applications. In summary, the contributions of this paper are as follows:

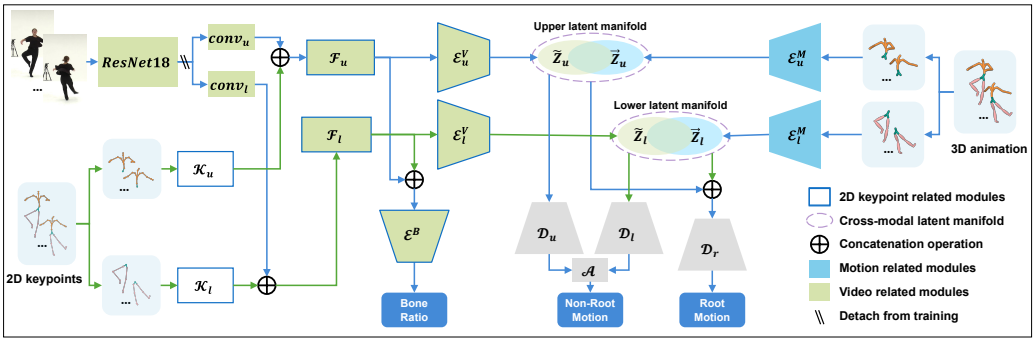


Figure 2: System overview of our VTM. The TPMAE (including motion encoders $\mathcal{E}_u^M$ & $\mathcal{E}_l^M$, motion decoders $\mathcal{D}_u$ & $\mathcal{D}_l$, and root decoder $\mathcal{D}_r$) are first trained on the motion data to learn two latent manifolds as the motion priors. Then, the TPVE (including 2D keypoints feature extractors $\mathcal{K}_u$ & $\mathcal{K}_l$, visual fusion blocks $\mathcal{F}_u$ & $\mathcal{F}_l$, visual encoders $\mathcal{E}_u^V$ & $\mathcal{E}_l^V$ and bone ratio predictor $\mathcal{E}^B$) are jointly trained with the pre-trained TPMAE to align the visual features with the motion priors for reconstructing 3D human motion. The superscripts $M$ and $V$ represent the "Motion" and "Video"; the subscripts $u$, $l$ and $r$ mean the "upper body part", "lower body part" and "root".

- We introduce a novel neural network, VTM, for learning human motion from video sequences. Our experimental results verify VTM's ability to faithfully replicate motion recorded in the provided videos.
- We present a new approach for harnessing scale-independent 3D motion priors by aligning motion data and video data on the two-body-part latent feature manifolds. This approach has proven effective in generating high-quality motion in our experiments.
- VTM produces a complete motion representation consisting of a kinematic skeleton, reliable global root translations, and naturally accurate joint rotations, thereby enabling more versatile applications in character motion and motion analysis.

## 2 Related Works

Human Pose Estimation (HPE) is a fundamental challenge in computer vision that plays a crucial role in various advanced applications. In simple terms, it involves the precise determination of the position and orientation of pivotal anatomical keypoints, such as the head, left hand, right foot, etc. when given a video, a signal, or an image. With the rapid growth of deep learning techniques, there has been a wealth of methods developed for 2D HPE in Andriluka et al. (2014); Chen et al. (2022); Munea et al. (2020); Dang et al. (2019); Zheng et al. (2023) and 3D HPE in Sarafianos et al. (2016); Wang et al. (2021); Chen et al. (2020). Furthermore, an array of methodologies pertaining to human body representation including SCAPE Anguelov et al. (2005), SMPL Loper et al. (2015), skeleton-based model Cao et al. (2018), surface-based model Güler et al. (2018), and Pose as Compositional Tokens (PCT) Geng et al. (2023b), as well as accompanying datasets, including HumanEva-I Sigal et al. (2010), Human3.6M Ionescu et al. (2013), MARCOnI Elhayek et al. (2016), Microsoft COCO Lin et al. (2014), etc.. have been proposed aiming to advance the field. In this section, we delve into three main paradigms within the forefront of 3D human pose estimation techniques: direct 3D HPE in Sec. 2.1, lifting from 2D to 3D pose in Sec. 2.2, and optimization & regression-based methods in Sec. 2.3.

### 2.1 Direct 3D HPE Methods

Similar to the approach employed in 2D HPE, the most straightforward method for 3D human pose estimation involves the utilization of an end-to-end neural network to make predictions regarding the 3D joint coordinates. Within this context, detection-based methods Pavlakos et al. (2017); Luvizon et al. (2018) predict a likelihood heatmap to determine the location of each joint, whilst regression-based methods treat the estimation of the joint locations relative to the root joint as a regression problem Li and Chan (2015); Zhou et al. (2016); Sun et al. (2017); Luvizon et al. (2019); Nibali

et al. (2018). In a notable development, Sun et al. Sun et al. (2018) innovatively integrated the regression approach in the joint position estimated from the heatmap, which effectively reduces the computational and storage cost associated with the end-to-end training process.

### 2.2 Lifting from 2D to 3D Pose

Given the unequivocal one-to-one correspondence between 2D and 3D joints, 3D HPE could benefit from the 2D HPE results as a means to enhance its in-the-wild generalization performance. For instance, utilizing a simple neural network to learn the lifting from 2D to 3D pose, which was popularized by the work in Martinez et al. (2017). These works Park et al. (2016); Tekin et al. (2017); Habibie et al. (2019); Zhou et al. (2019a) have made significant progress in addressing the inherent challenge by tackling the fusion of 2D heatmaps with 3D image cues. Other methods, including long short-term memory (LSTM) Nie et al. (2019); Wang et al. (2019), Euclidean distance matrix Moreno-Noguer (2017), and graph neural networks Zhao et al. (2019); Ci et al. (2019), have leveraged the correspondence between joints for designing advanced algorithms. During training, it is often to employ 2D space reprojections of the 3D pose as a form of supervision to ensure consistency Tome et al. (2017); Habibie et al. (2019); Chen et al. (2019). The Transformer model Vaswani et al. (2017) additionally furnishes an alternative avenue for optimizing the transformation process from 2D representations to their corresponding 3D counterparts Einfalt et al. (2023); Li et al. (2022a,b); Shan et al. (2022); Zhang et al. (2022); Tang et al. (2023). Zheng et al. Zheng et al. (2021) have proposed PoseFormer, which utilizes cascaded transformer layers to achieve remarkable performance, and PoseFormerV2 Zhao et al. (2023), where enhancements in both efficiency and robustness have been achieved. Furthermore, with the introduction of GANs(Generative Adversarial Networks), more realistic 3D human poses have been generated with the help of a discriminator Fish Tung et al. (2017); Yang et al. (2018); Wandt and Rosenhahn (2019).

### 2.3 Optimization and Regression-based Methods

In recent times, the skinned multi-person linear (SMPL) model Loper et al. (2015) has become a popular choice for the representation of human bodies when fitting to images employing annotated keypoints and silhouettes. Many optimization techniques tailored for the SMPL model were introduced for 3D HPE Bogo et al. (2016); Lassner et al. (2017); Han et al. (2023); Liang et al. (2023) due to its inherent capacity to incorporate prior knowledge concerning human body shape, rendering it amenable to effective fitting even with limited training data. However, optimization-based approaches relying solely on these fitting techniques may lead to unnatural body shapes and poses. Some other methods regress the SMPL parameters by various networks Kanazawa et al. (2018); Pavlakos et al. (2018); Omran et al. (2018); Zhang et al. (2021); Sun et al. (2021, 2022) due to their effectiveness, whilst regression-based approaches encounter challenges in maintaining precise alignment between images and the corresponding mesh representations. Moreover, the Inverse Kinematics Optimization Layer (IKOL) Zhang et al. (2023a) utilizes the SPIN (SMPL oPtimization IN the loop) Kolotouros et al. (2019) framework to amalgamate optimization-based and regression-based methodologies, whose synergy leads to a marked improvement in performance.

In this paper, we focus on reconstructing the skeleton, the joint rotations, and the global root translations from the video features and 2D poses by aligning them with the pre-learned motion priors.

## 3 Video-to-Motion Generator

### 3.1 Data Preparation

**Dataset.** A cross-modal dataset containing both 3D joint rotation data and 2D videos, like the AIST++ Li et al. (2021b) and Human3.6M Ionescu et al. (2013) dataset, is indispensable for our objective of generating a complete motion representation for diverse characters from monocular videos. The AIST++ dataset contains far more performers than the Human3.6M dataset (27 vs. 7), which means there are more skeleton variations in AIST++. This makes it more challenging to model the motion in AIST++ but more suitable for our task since we want to reconstruct the skeletons from the 2D inputs. Therefore, we opt for the AIST++ dataset to train and evaluate our VTM. Our

experiments only utilize the 2D keypoints and videos captured by camera 1 as our single-view 2D input; more details about the camera settings can be found in Tsuchida et al. (2019).

**Data Preprocess.** Firstly, we inspect the dataset, excluding sequences that exhibit incorrect poses, to ensure the learning of a well-defined latent manifold. Subsequently, leveraging the 3D joint positions, we construct $J$-joint skeletons, adjusting bone lengths to match the characters in the videos. Finally, employing the joint rotations and the created skeletons, we generate motion files in BVH format, resulting in the BVHAIST++ dataset. This dataset will be publicly available alongside our source code.

Following this, we construct a virtual skeleton, denoted as $\bar{\mathbf{s}}$, by averaging the skeletons present in BVHAIST++. All motion data undergo alignment with $\bar{\mathbf{s}}$, yielding a skeleton-scale-independent motion dataset. This dataset facilitates the learning of motion priors disentangled from the specifics of the skeleton scale. Additionally, we define the bone ratios $\mathbf{b}$ as the ratios between the original skeletons and the average skeleton. During inference, the ratios predicted by VTM are used to recover the original skeletons from the averaged skeleton.

**3D Motion Representation.** We convert the BVH files of $\bar{\mathbf{s}}$ into camera space using the camera parameters and represent the $t_{th}$ frame of the transformed motion sequence as $\mathbf{x}_t = \{\mathbf{r}_t^q, \mathbf{r}_t^p, \mathbf{r}_t^v, \mathbf{x}_t^q, \mathbf{x}_t^p, \mathbf{x}_t^v\}$. Here, $\mathbf{r}_t^q \in \mathbb{R}^6$ and $\mathbf{x}_t^q \in \mathbb{R}^{6(J-1)}$ denote the 6D rotation representations Zhou et al. (2019b) of the root and non-root joints. $\mathbf{r}_t^p \in \mathbb{R}^3$ and $\mathbf{x}_t^p \in \mathbb{R}^{3(J-1)}$ represent the global 3D joint positions, while $\mathbf{r}_t^v \in \mathbb{R}^3$ and $\mathbf{x}_t^v \in \mathbb{R}^{3(J-1)}$ are the linear velocities.

**2D Input Representation.** To maintain scale consistency between 2D keypoints and 3D motion, we project the 3D joint positions obtained through the forward kinematic (FK) process into 2D utilizing the camera parameters. These 2D keypoints are referred to as virtual 2D keypoints. Then, we represent the $t_{th}$ frame of the 2D keypoints input as $\mathbf{k}_t = \{\mathbf{k}_t^p, \mathbf{k}_t^v\}$, where $\mathbf{k}_t^p \in \mathbb{R}^{2J}$ represents the 2D virtual keypoints, and $\mathbf{k}_t^v \in \mathbb{R}^{2J}$ corresponds to their linear velocities.

### 3.2 Learn Motion Priors

To learn compact and well-defined priors from the motion data, we design a two-part motion autoencoder (TPMAE) comprised of the motion encoders, motion decoders, and root decoder depicted in Fig. 2. TPMAE integrates the two-part approach, inspired by the works of Hou et al. (2023a); Tao et al. (2023), into a convolutional architecture closely resembling the one employed in Zhang et al. (2023b) (the details can be found in the supplementary material).

**Encoders.** Given an input sequence comprising $T$ frames, denoted as $\mathbf{X} = \{\mathbf{x}_t, ..., \mathbf{x}_{t+T-1}\} \in \mathbb{R}^{T\times J\times 12}$, we first partition it into two sub-sequences based on joint indices: $\mathbf{X}_u$ for the upper body part and $\mathbf{X}_l$ for the lower body part, both contain root data. We then employ two motion encoders, namely $\mathcal{E}_u^M$ and $\mathcal{E}_l^M$, to model the two sub-sequences. This partitioning strategy can significantly reduce the complexity of modeling the entire human pose manifold, as stated in Hou et al. (2023a). The encoders share a similar architecture to the encoder introduced by Zhang et al. Zhang et al. (2023b). As a result, the latent vectors $\mathbf{Z}_u \in \mathbb{R}^{\frac{T}{4}\times 128}$ and $\mathbf{Z}_l \in \mathbb{R}^{\frac{T}{4}\times 64}$ for the upper and lower bodies are formulated as follows:

$$\mathbf{Z}_u = \mathcal{E}_u^M(\mathbf{X}_u), \mathbf{Z}_l = \mathcal{E}_l^M(\mathbf{X}_l) \tag{1}$$

We treat $\mathbf{Z}_u$ and $\mathbf{Z}_l$ as our motion priors. Note that the input $\mathbf{X}$ has been aligned with the virtual skeleton $\bar{\mathbf{s}}$. Consequently, the encoders are capable of capturing 3D motion kinematic constraints that are independent of skeleton scales, without the interference of skeleton-scale variations. As detailed in Sec. 4.3.1, these skeleton-scale-independent motion priors contribute to a reduced MPJPE in the reconstructed motion.

**Decoders.** On the latent manifold, two decoders, $\mathcal{D}_u$ and $\mathcal{D}_l$, are tasked with decoding $\mathbf{Z}_u$ and $\mathbf{Z}_l$, respectively. Additionally, an aggregation layer $\mathcal{A}$, formed by a 1D convolution layer, is exploited to map the decoded features of body parts to the non-root motion $\hat{\mathbf{X}}_{nr}$. Simultaneously, another decoder, $\mathcal{D}_r$, is responsible for directly decoding the concatenation of $\mathbf{Z}_u$ and $\mathbf{Z}_l$ back into the root motion $\hat{\mathbf{X}}_r$. This approach of independently reconstructing root and non-root joints has been proven effective by Li et al. Siyao et al. (2022). In short, the decoding process can be represented as follows:

$$\hat{\mathbf{X}}_{nr} = \mathcal{A}(\mathcal{D}_u(\mathbf{Z}_u) \oplus \mathcal{D}_l(\mathbf{Z}_l)), \hat{\mathbf{X}}_r = \mathcal{D}_r(\mathbf{Z}_u \oplus \mathbf{Z}_l) \tag{2}$$

where $\oplus$ represents the concatenation operation. Note that $\hat{\mathbf{X}}_r$ only contains the root rotations $\mathbf{r}^q$, the 3D root positions $rz^p$ on the Z-axis and the corresponding velocities $rz^v$. During the inference, the global root translations can be derived from $rz^p$, the coordinates of the given 2D keypoints, and the intrinsic parameters of the camera model.

**Training Losses.** The loss function $L^M$ used to train TPMAE contains two terms: the motion reconstruction loss $L^M_{rec}$ and the motion smoothness loss $L^M_s$, as follows:

$$L^M = L^M_{rec} + L^M_s \tag{3}$$

*Motion reconstruction loss* is used to constrain the latent manifold to be well-defined. Since different joints matter differently in the motion, we scale the reconstructed and ground-truth motion data by giving relative importance $\boldsymbol{\omega}_r$ and $\boldsymbol{\omega}_{nr}$ to the root joint and other joints like Holden et al. Holden et al. (2017) did. Therefore, $L^M_{rec}$ can be computed as:

$$L^M_{rec} = L_1(\boldsymbol{\omega}_r\hat{\mathbf{X}}_r, \boldsymbol{\omega}_r\mathbf{X}_r) + L_1(\boldsymbol{\omega}_{nr}\hat{\mathbf{X}}_{nr}, \boldsymbol{\omega}_{nr}\mathbf{X}_{nr}) \tag{4}$$

where $L_1$ represents the smooth L1 loss function. Specifically, we set the relative importance for the root to be 2.0, the end-effectors to be 1.5, and all other joints to be 1.0.

*Smoothness loss* plays a vital role in preventing the reconstructed motion from abrupt variations along the temporal axis. Furthermore, it bestows added significance to the root, resulting in the following formulation:

$$\begin{aligned} L^M_s =& L_1(\boldsymbol{\omega}_r\hat{\mathbf{Vel}}_r, \boldsymbol{\omega}_r\mathbf{Vel}_r) + L_1(\hat{\mathbf{Vel}}_{nr}, \mathbf{Vel}_{nr}) + \\ & L_1(\boldsymbol{\omega}_r\hat{\mathbf{Acc}}_r, \boldsymbol{\omega}_r\mathbf{Acc}_r) + L_1(\hat{\mathbf{Acc}}_{nr}, \mathbf{Acc}_{nr}) \end{aligned} \tag{5}$$

where $\mathbf{Vel}$ & $\hat{\mathbf{Vel}}$ and $\mathbf{Acc}$ & $\hat{\mathbf{Acc}}$ represent the velocities and accelerations of the ground-truth & reconstructed data. Take the $t_{th}$ frame for example, its velocity $\mathbf{vel}_t = \mathbf{x}_t - \mathbf{x}_{t-1}$, and acceleration $\mathbf{acc}_t = \mathbf{vel}_t - \mathbf{vel}_{t-1}$.

### 3.3 Learn Human Motion from Videos

After learning the motion priors, we introduce a two-part visual encoder (TPVE) to map the visual features into the learned motion priors. As illustrated in Fig. 2, the TPVE consists of a video feature extractor, two 2D keypoints feature extractors for the upper and lower body parts, two visual feature fusion blocks to fuse the video features and keypoints features, two streams of visual encoders to map the fused visual features into the latent manifold, and a bone ratio predictor to predict the bone scale.

**Video Feature Extractor.** We utilize the pre-trained ResNet18 He et al. (2016) as our video feature extractor, where the last fully connected layer is discarded. The weights of the pre-trained ResNet18 are fixed during the training procedure. We also employ two learnable 1D convolutional layers, $conv_u$, and $conv_l$, to facilitate our TPVE with the ability to adjust the video features.

**2D Keypoints Feature Extractors.** Similar to the process of motion data, we first divide the 2D keypoint sequences $\mathbf{K} \in \mathbb{R}^{T\times J\times 4}$ into two sub-sequences: $\mathbf{K}_u$ for the upper body part and $\mathbf{K}_l$ for the lower body part. Subsequently, we employ two three-layer CNNs, $\mathcal{K}_u$ and $\mathcal{K}_l$, for these two body parts to extract the 2D keypoint features.

**Visual Feature Fusion Blocks.** After obtaining the video features and keypoints features, we adopt two residual blocks $\mathcal{F}_u$ and $\mathcal{F}_l$ structured with 1D convolution layers to fuse them, producing the visual features $\tilde{\mathbf{V}}_u$ and $\tilde{\mathbf{V}}_l$:

$$\begin{aligned} \tilde{\mathbf{V}}_u &= \mathcal{F}_u(conv_u(\mathbf{V}) \oplus \mathcal{K}_u(\mathbf{K}_u)) \\ \tilde{\mathbf{V}}_l &= \mathcal{F}_l(conv_l(\mathbf{V}) \oplus \mathcal{K}_l(\mathbf{K}_l)) \end{aligned} \tag{6}$$

**Visual Encoders.** When we obtain the visual features, we utilize two encoders, $\mathcal{E}^V_u$ and $\mathcal{E}^V_l$, for the upper and lower body parts, to map these visual features into the latent manifold. These encoders have structures similar to the motion encoders but followed by two cross-temporal context aggregation modules (CTCA) Guo et al. (2023), to better capture the temporal correlations. Then, the latent manifold can be encoded by:

$$\tilde{\mathbf{Z}}_u = \mathcal{E}^V_u(\tilde{\mathbf{V}}_u), \tilde{\mathbf{Z}}_l = \mathcal{E}^V_l(\tilde{\mathbf{V}}_l) \tag{7}$$

where $\tilde{\mathbf{Z}}_u \in \mathbb{R}^{\frac{T}{4}\times 128}$ and $\tilde{\mathbf{Z}}_l \in \mathbb{R}^{\frac{T}{4}\times 64}$ represent the latent manifold encoded from the visual features.

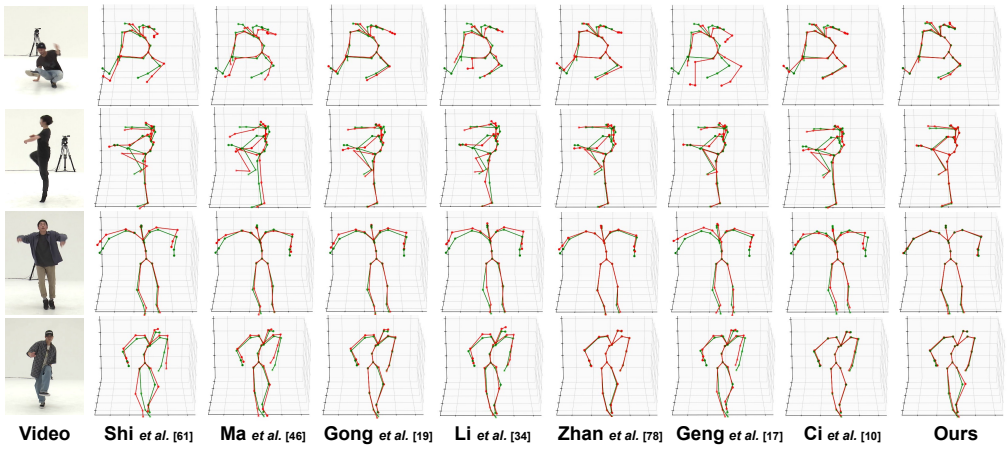


Figure 3: Qualitative comparisons to other SOTA methods. The green skeletons represent the ground truth poses, and the red ones represent the reconstructed poses by different methods.

**Bone Ratios Predictor.** The scale of the skeleton plays a crucial role in 3D human motion due to substantial variations in human height and proportions for different body parts. While we deliberately remove scale information from the training motion data to enhance the learning of motion priors, we can effectively extract bone ratios $\mathbf{b}$ from the 2D inputs. These inputs inherently embody the scale information of the character. To accomplish this, we utilize another encoder $\mathcal{E}^B$ bearing a similar structure as the motion encoders to predict the bone ratio as follows:

$$\tilde{\mathbf{b}} = \mathcal{E}^B(\tilde{\mathbf{V}}_u \oplus \tilde{\mathbf{V}}_l) \tag{8}$$

during inference, the skeleton can be reconstructed by $\tilde{\mathbf{b}} * \bar{s}$. **Training Losses.** To empower the TPVE with the capability to encode visual features into motion priors, we employ a manifold alignment loss $L_{ma}$ and a bone prediction loss $L_b$ along with the motion prediction and smoothness loss. The TPVE is jointly trained with the TPMAE, resulting in the final loss function of $L$ as follows:

$$L = L_{ma} + L_b + L^V_{pred} + L^V_s + L^M \tag{9}$$

*Manifold alignment loss* aims to align the visual latent manifold with the motion priors. To achieve this, we use the smooth L1 loss which is calculated as follows:

$$L_{ma} = L_1(\tilde{\mathbf{Z}}_u, \mathbf{Z}_u) + L_1(\tilde{\mathbf{Z}}_l, \mathbf{Z}_l) \tag{10}$$

Although this method is simple, it surprisingly performs well in aligning the cross-modal latent manifolds for our task. We will further discuss its efficacy in Sec. 4.3.2.

*Bone prediction loss* is a smooth L1 loss as follows:

$$L_b = L_1(\hat{\mathbf{b}}, \mathbf{b}) \tag{11}$$

*Motion prediction and smoothness loss* are the same as those employed in TPMAE, which ensures the visual latent vectors can be decoded into the motion data. They can be computed as follows:

$$L^V_{pred} = L_1(\boldsymbol{\omega}_r \tilde{\mathbf{X}}_r, \boldsymbol{\omega}_r \mathbf{X}_r) + L_1(\boldsymbol{\omega}_{nr} \tilde{\mathbf{X}}_{nr}, \boldsymbol{\omega}_{nr} \mathbf{X}_{nr}) \tag{12}$$

$$\begin{aligned} L^V_s =& L_1(\boldsymbol{\omega}_r \tilde{\mathbf{Vel}}_r, \boldsymbol{\omega}_r \mathbf{Vel}_r) + L_1(\tilde{\mathbf{Vel}}_{nr}, \mathbf{Vel}_{nr}) + \\ & L_1(\boldsymbol{\omega}_r \tilde{\mathbf{Acc}}_r, \boldsymbol{\omega}_r \mathbf{Acc}_r) + L_1(\tilde{\mathbf{Acc}}_{nr}, \mathbf{Acc}_{nr}) \end{aligned} \tag{13}$$

where $\tilde{\mathbf{X}}_r$ and $\tilde{\mathbf{X}}_{nr}$ are the root and non-root data reconstructed from the visual inputs by the decoders. Similarly, $\tilde{\mathbf{Vel}}$ and $\tilde{\mathbf{Acc}}$ represent the velocities and accelerations of the predicted motion.

After training, the motion encoders are discarded, and we only feed the VTM with visual features to generate human motion. The detailed architectures and parameters of all the modules of TPMAE and TPVE will be reported in the supplementary material.

| Method | MPJPE↓ | PA-MPJPE↓ | MRPE↓ | MBLE↓ |
|---|---|---|---|---|
| Shi et al. (2020) | 37.0 | 33.4 | 45.3 | 1.4 |
| Ma et al. (2021) | 20.1 | 14.8 | - | 10.4 |
| Gong et al. (2021) | 24.8 | 19.6 | - | 4.8 |
| Li et al. (2021a) | 45.1 | 37.4 | - | 9.5 |
| Zhan et al. (2022) | 13.2 | 10.8 | 141.4 | 3.5 |
| Geng et al. (2023a) | 23.7 | 17.8 | - | 5.4 |
| Ci et al. (2023) | 27.3 | 19.2 | - | 6.0 |
| Ours | 17.8 | 15.7 | 16.8 | 0.1 |

Table 1: Comparisons to other SOTA 3D HPE methods on MPJPE, PA-MPJPE, MRPE, and MBLE. The best results are highlighted as 1st, 2nd, and 3rd.

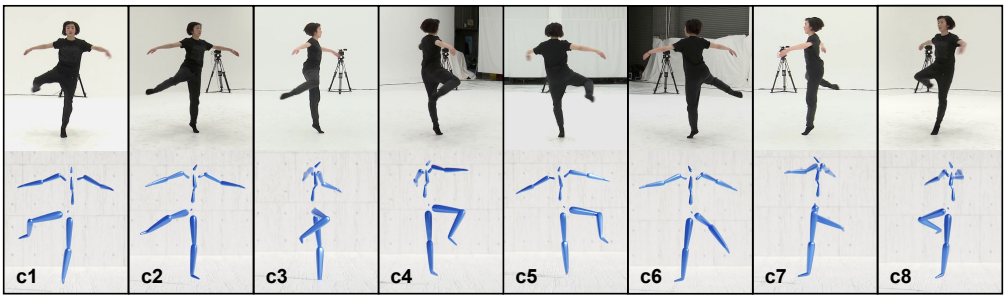


Figure 4: VTM can generalize to different view angles. The first row is the video frames from different camera settings, and the second row is the same pose viewed from the angles corresponding to the camera settings. Only videos and 2D keypoints from camera setting 1, namely c1, are used for training our VTM.

# 4 Experiments

## 4.1 Implementation Details

We implement our algorithm using PyTorch 1.10.1 on a desktop PC equipped with two GeForce RTX 3090 24GB graphics cards and an Intel(R) Xeon(R) E5-2678 CPU with 128GB RAM.

In our experimental setup, the joint number $J$ of the skeleton is set to be 24. The training sequences for the VTM consist of 32 frames each, consecutively extracted from the dataset using a sliding window with a length of 4. For TPMAE, we employ the AdamW optimizer Loshchilov and Hutter (2017) for training, running for 500 epochs with a batch size of 100. The initial learning rate is set at 1e-4, subject to decay by multiplication with 0.5 every 100 epochs. Subsequently, the TPVE undergoes joint training with TPMAE using the same settings, except for a reduced batch size of 64.

## 4.2 Comparisons

To the best of our knowledge, MotioNet Shi et al. (2020) is the only existing method that produces a complete 3D human motion representation from 2D input. For a more comprehensive evaluation, we also compare VTM with other SOTA 3D HPE methods that employ different model structures and/or inputs only to predict 3D joint positions. For fair comparisons, we utilize these methods' released codes and training settings to train their models on the AIST++ dataset. Note that the input 2D keypoints for these methods are projected from the corresponding 3D joint positions. Please refer to the supplementary material for the detailed settings of these methods.

**Quantitative comparison.** We employ MPJPE (without rigid alignment) and PA-MPJPE (with rigid alignment), following the common practice in this field, for a fair comparison of joint position reconstruction accuracy. We also employ the mean of 3D bone length error (MBLE) to evaluate the benefits of reconstructing the complete motion representation. Additionally, we assess the mean root position error (MRPE) in comparison to MotioNet and Ray3D Zhan et al. (2022), as these two

| Method | MPJPE↓ | PA-MPJPE↓ | MRPE↓ |
|---|---|---|---|
| TPMAE-SKW | 4.9 | 4.1 | 0.8 |
| TPMAE-Q | 84.0 | 10.7 | 98.5 |
| OPMAE | 5.7 | 4.9 | 1.1 |
| TPMAE | 4.8 | 4.0 | 0.8 |

Table 2: Evaluations of TPMAE on MPJPE, PA-MPJPE, and MRPE.

methods reconstruct global root translations as well. All metrics are in millimeters and computed on the validation dataset .

Since MotioNet and VTM generate the joint rotations, we perform FK using the reconstructed skeletons and joint rotations to obtain 3D joint positions for metric computation. In Tab. 1, VTM ranks second and third in terms of MPJPE and PA-MPJPE. This can be attributed to VTM's slightly different objective, focusing on reconstructing a complete 3D motion representation from monocular videos, setting it apart from methods primarily concentrating on 3D joint positions. Notably, VTM excels in MRPE compared to MotioNet and Ray3D, benefiting from encoding global root positions into motion priors along with the joint rotations. Additionally, VTM and MotioNet secure the top two positions in MBLE, showcasing the effectiveness of reconstructing complete motion representation in mitigating bone length inconsistency issues encountered when reconstructing 3D joint positions.

We train VTM on the Human3.6M dataset and achieve the three metrics on the test dataset (S9 and S11) as (67.3, 49.1, 15.6). It's worth noting that the Human3.6M skeleton includes three additional joints corresponding to zero-length bones, which are crucial for skeleton-and-rotation-driven human motion. However, these joints may bring complexity during network optimization. If our focus is solely on 3D joint positions, the issue of rotation confusion becomes less relevant. This contextualizes why VTM's MPJPE metrics might not match those of SOTA methods like Ray3D, but the MRPE surpasses Ray3D's reported value of 109.5mm.

**Qualitative comparison.** To visually compare the quality of reconstructed motion among different methods, we randomly select four video frames from the validation set and show the reconstructed 3D joint positions in Fig. 3. Notably, VTM consistently surpasses or is on par with other methods. Despite quantitative and qualitative weaknesses compared to Ray3D in many cases, VTM exhibits superiority when significant self-occlusions occur in the video, exemplified in the second row of Fig. 3. We attribute this advantage to our motion priors, which encode reasonable 3D motion manifolds. This effective encoding captures correlations between joint rotations, enabling the inference of plausible human body poses in such situations.

### 4.3 Evaluations

#### 4.3.1 Evaluations of TPMAE

We conduct experiments to validate our TPMAE's efficacy through various configurations: 1) TPMAE-SKW: In this scenario, TPMAE is trained using motion data in original skeletons instead of those aligned with $\bar{s}$. This investigation aims to discern the impact of skeleton scales on motion reconstruction accuracy. 2) TPMAE-Q: Using quaternion-based rotation and a local 3D joint position representation similar to that used in Holden et al. (2017); Hou et al. (2023a). 3) OPMAE: To assess the necessity of the two-part design in TPMAE, we discard the two-part structure and increase the latent space dimension to 196. This adjustment is made to keep the size of the latent manifold unchanged.

We perform forward kinematics using the predicted joint rotations and ground-truth skeletons, presenting the quantitative results in Tab. 2. The superiority of TPMAE over its variants on these metrics confirms its effectiveness. In TPMAE-Q, representing root rotations as angular velocities around the Y-axis leads to the accumulation of errors, especially in long sequences. This explains why TPMAE-Q performs comparably on PA-MPJPE but poorly on MPJPE compared to other models.

#### 4.3.2 Evaluations of VTM

**Generalization to unseen view angles.** As Sec. 3.1 mentions, our VTM is exclusively trained using the 2D keypoints and videos obtained from camera 1. In this particular experiment, we evaluate

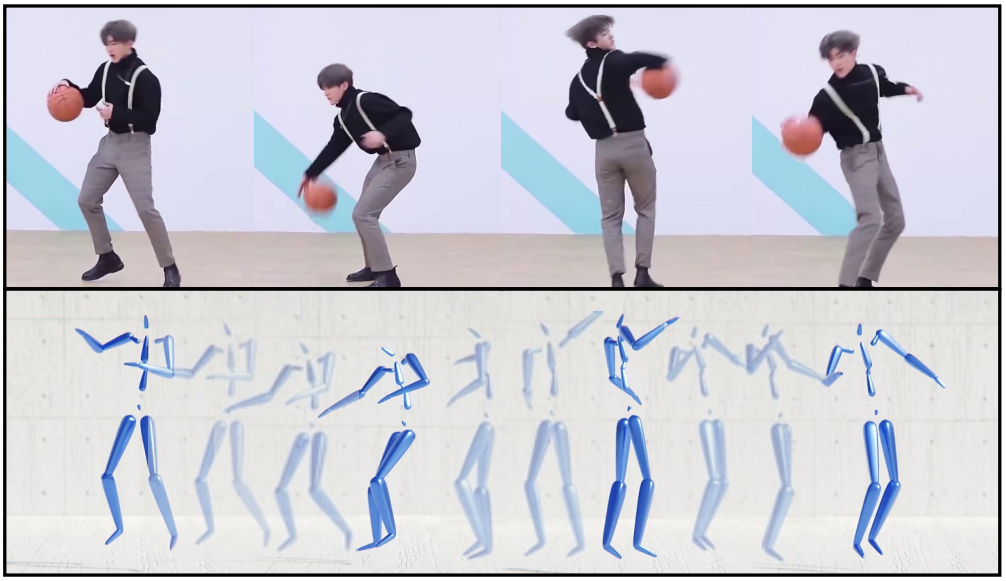

Figure 5: VTM can reconstruct motion from in-the-wild videos. The first row is the continuous frames from a wild video, and the second row shows the corresponding results produced by VTM.

| Method | MPJPE↓ | PA-MPJPE↓ | MRPE↓ |
|---|---|---|---|
| VTM-CL | 22.5 | 19.4 | 21.1 |
| VTM-CL+L1 | 20.5 | 17.8 | 19.2 |
| VTM-SKW | 19.1 | 16.8 | 22.2 |
| VTM-OP | 72.1 | 58.8 | 65.2 |
| VTM-w/o_JT | 21.8 | 18.1 | 24.3 |
| VTM-w/o_PT | 20.9 | 18.5 | 21.3 |
| VTM-w/o_CTCA | 21.0 | 18.4 | 20.8 |
| VTM-w/o_K | 73.5 | 59.7 | 66.5 |
| VTM-w/o_V | 20.2 | 17.5 | 31.8 |
| VTM | 17.8 | 15.7 | 16.8 |

Table 3: Evaluations of VTM on MPJPE, PA-MPJPE, and MRPE. The best results are highlighted as 1st, 2nd, and 3rd.

the performance of our VTM by utilizing the 2D data collected from cameras 2 to 8 in the AIST++ dataset. As illustrated in Fig. 4, our VTM can generalize well across previously unseen view angles.

**Generalization to in-the-wild videos.** We introduce a mapper network to enable VTM to reconstruct motion from in-the-wild videos. This network, constructed as a four-layer MLP, is designed to convert COCO-formatted keypoints to our virtual 2D keypoints. This mapper facilitates the seamless application of our VTM to in-the-wild videos, so that we can directly utilize 2D keypoints detected by readily available 2D HPE models. As illustrated in Fig. 5 and e) & f) of Fig. 1, the VTM consistently produces plausible 3D human motion from in-the-wild videos.

**Evaluations on motion priors.** Our VTM leverages pre-learned motion priors by aligning the visual latent manifold with them, undergoing joint training with the pre-trained TPMAE. To assess its effectiveness, we conducted the following experiments.

*VTM-CL and VTM-CL+L1.* These experiments aim to identify a suitable method for aligning cross-modal latent manifolds. In VTM-CL, we replaced $L_{ma}$ with a contrastive loss (CL), similar to the methodology used in CLIP. This loss function is calculated based on scaled pairwise cosine similarities between visual latent vectors and motion prior vectors. In VTM-CL+L1, we integrated the CL loss into $L_{ma}$. However, as indicated in Tab. 3, both variants are inferior to our VTM. We posit that the CL loss may be unsuitable for our task due to the presence of numerous similar poses

in the dataset. Selecting training batches in a sliding window manner can introduce sequences with higher similarity. Simply treating such similar sequences in a batch as negative samples may confuse the optimization of the network.

*VTM-w/o_JT and VTM-w/o_PT.* To explore joint training with the pre-trained TPMAE, we conduct two additional experiments: 1) training only the TPVE while keeping the weights of TPMAE decoders fixed (VTM-w/o_JT), and 2) directly training TPMAE and TPVE end-to-end without pre-training TPMAE (VTM-w/o_PT). As demonstrated in Tab. 3, the joint training strategy can significantly improve the accuracy of motion and global root translation reconstruction.

**Other Evaluations.** The efficacy of various configurations of our VTM is assessed through a series of experiments, with quantitative results reported in Tab. 3.

*VTM-SKW.* Utilizing TPMAE-SKW as the motion auto-encoder, this model is trained on motion data in original skeletons to examine the impact of skeleton scales on 3D motion reconstruction. The results demonstrate that aligning human motion to our virtual skeleton significantly enhances motion reconstruction from monocular videos. By excluding skeleton scale information from the training data, VTM focuses solely on learning kinematic constraints such as joint rotations, which proves advantageous for our task.

*VTM-OP.* We abandon the two-part strategy and use the OPMAE as the motion auto-encoder to analyze the advantages of our two-part design. Despite the comparable performance of OPMAE to TPMAE, the VTM-OP lags behind VTM by a large margin, indicating that learning human motion from monocular videos is far more challenging than motion reconstruction.

*VTM-w/o_CTCA.* This model is trained with the exclusion of the CTCA modules from the visual encoders. Tab. 3 demonstrates that CTCA plays a crucial role in enhancing the temporal correlations captured by VTM, resulting in higher motion reconstruction accuracy.

*VTM-w/o_K and VTM-w/o_V.* We investigate how different 2D inputs influence VTM performance. When fed only with videos (VTM-w/o_K), there is a significant decrease in performance across all metrics. Conversely, when provided with only 2D keypoints (VTM-w/o_V), VTM-w/o_V maintains top-3 performance in MPJPE but experiences a notable drop in MRPE. This suggests that 2D keypoints assist the model in identifying motion contours, while video information contributes to global spatial awareness.

## 5 Conclusion

In conclusion, we present an innovative framework named VTM, specializing in learning human motion from monocular videos. Our VTM first learns motion priors from 3D motion data and subsequently leverages these well-defined priors to reconstruct high-fidelity motion from 2D inputs. We conduct comprehensive experiments to validate the effectiveness of our proposed method. Notably, VTM exhibits satisfactory generalization capabilities across various unseen view angles and performs robustly on in-the-wild videos.

**Future work.** Most high-quality human motion datasets acquired by professional motion capture systems, such as CMU mocap CMU (2023), are unfortunately devoid of paired video data, rendering them unsuitable for our current supervised learning approach. In the future, we plan to investigate unsupervised or semi-supervised learning methods. This exploration seeks to leverage existing high-quality human motion datasets alongside readily available video data, intending to enhance the model's generalization capabilities and robustness.

## References

Ijaz Akhter and Michael J. Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1446–1455, 2015.

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3686–3693, 2014.

Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *CoRR*, abs/1812.08008, 2018.

Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Rohith Mv, Stefan Stojanov, and James M Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5714–5724, 2019.

Haoming Chen, Runyang Feng, Sifan Wu, Hao Xu, Fengcheng Zhou, and Zhenguang Liu. 2d human pose estimation: A survey. *Multimedia Systems*, pages 1–24, 2022.

Yucheng Chen, Yingli Tian, and Mingyi He. Monocular human pose estimation: A survey of deep learning-based methods. *Computer vision and image understanding*, 192:102897, 2020.

Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *Int. Conf. Comput. Vis.*, pages 2262–2271, 2019.

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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4800–4810, 2023.

CMU. Cmu graphics lab motion capture database, 2023.

Qi Dang, Jianqin Yin, Bin Wang, and Wenqing Zheng. Deep learning based 2d human pose estimation: A survey. *Tsinghua Science and Technology*, 24(6):663–676, 2019.

Andrey Davydov, Anastasia Remizova, Victor Constantin, Sina Honari, Mathieu Salzmann, and Pascal Fua. Adversarial parametric pose prior. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10987–10995, 2022.

Moritz Einfalt, Katja Ludwig, and Rainer Lienhart. Uplift and upsample: Efficient 3d human pose estimation with uplifting transformers. In *IEEE Winter Conf. Appl. Comput. Vis.*, pages 2903–2913, 2023.

Ahmed Elhayek, Edilson de Aguiar, Arjun Jain, Jonathan Thompson, Leonid Pishchulin, Mykhaylo Andriluka, Christoph Bregler, Bernt Schiele, and Christian Theobalt. Marconi—convnet-based marker-less motion capture in outdoor and indoor scenes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(3):501–514, 2016.

Hsiao-Yu Fish Tung, Adam W Harley, William Seto, and Katerina Fragkiadaki. Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision. In *Int. Conf. Comput. Vis.*, pages 4354–4362, 2017.

Zigang Geng, Chunyu Wang, Yixuan Wei, Ze Liu, Houqiang Li, and Han Hu. Human pose as compositional tokens. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023a.

Zigang Geng, Chunyu Wang, Yixuan Wei, Ze Liu, Houqiang Li, and Han Hu. Human pose as compositional tokens. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 660–671, 2023b.

Kehong Gong, Jianfeng Zhang, and Jiashi Feng. Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7297–7306, 2018.

Leming Guo, Wanli Xue, Qing Guo, Bo Liu, Kaihua Zhang, Tiantian Yuan, and Shengyong Chen. Distilling cross-temporal contexts for continuous sign language recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10771–10780, 2023.

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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10905–10914, 2019.

Xiao Han, Peishan Cong, Lan Xu, Jingya Wang, Jingyi Yu, and Yuexin Ma. Licamgait: Gait recognition in the wild by using lidar and camera multi-modal visual sensors. In *AAAI*, 2023.

H Hatze. A three-dimensional multivariate model of passive human joint torques and articular boundaries. *Clinical Biomechanics*, 12(2):128–135, 1997.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016.

Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Trans. Graph.*, 36(4):1–13, 2017.

Shuaiying Hou, Hongyu Tao, Hujun Bao, and Weiwei Xu. A two-part transformer network for controllable motion synthesis. *IEEE Trans. Vis. Comput. Graph.*, pages 1–16, 2023a.

Shuaiying Hou, Congyi Wang, Wenlin Zhuang, Yu Chen, Yangang Wang, Hujun Bao, Jinxiang Chai, and Weiwei Xu. A causal convolutional neural network for multi-subject motion modeling and generation. *Computational Visual Media*, 10(1):45–59, 2023b.

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 Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2013.

Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7122–7131, 2018.

T. Kodek and M. Munih. Identifying shoulder and elbow passive moments and muscle contributions during static flexion-extension movements in the sagittal plane. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1391–1396 vol.2, 2002.

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 *Int. Conf. Comput. Vis.*, pages 2252–2261, 2019.

Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6050–6059, 2017.

Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *Int. Conf. Comput. Vis.*, 2021a.

Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++, 2021b.

Sijin Li and Antoni B Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *ACCV*, pages 332–347. Springer, 2015.

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 Trans. Multimedia*, 25:1282–1293, 2022a.

Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13147–13156, 2022b.

Han Liang, Yannan He, Chengfeng Zhao, Mutian Li, Jingya Wang, Jingyi Yu, and Lan Xu. Hybridcap: Inertia-aid monocular capture of challenging human motions. In *AAAI*, pages 1539–1548, 2023.

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.

Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. Character controllers using motion vaes. *ACM Trans. Graph.*, 39(4), 2020.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):1–16, 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5137–5146, 2018.

Diogo C Luvizon, Hedi Tabia, and David Picard. Human pose regression by combining indirect part detection and contextual information. *Computers & Graphics*, 85:15–22, 2019.

Xiaoxuan Ma, Jiajun Su, Chunyu Wang, Hai Ci, and Yizhou Wang. Context modeling in 3d human pose estimation: A unified perspective. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6238–6247, 2021.

Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Int. Conf. Comput. Vis.*, pages 2640–2649, 2017.

Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2823–2832, 2017.

Tewodros Legesse Munea, Yalew Zelalem Jembre, Halefom Tekle Weldegebriel, Longbiao Chen, Chenxi Huang, and Chenhui Yang. The progress of human pose estimation: A survey and taxonomy of models applied in 2d human pose estimation. *IEEE Access*, 8:133330–133348, 2020.

Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast. Numerical coordinate regression with convolutional neural networks. *arXiv preprint arXiv:1801.07372*, 2018.

Xuecheng Nie, Jiashi Feng, Jianfeng Zhang, and Shuicheng Yan. Single-stage multi-person pose machines. In *Int. Conf. Comput. Vis.*, pages 6951–6960, 2019.

Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, pages 484–494. IEEE, 2018.

Sungheon Park, Jihye Hwang, and Nojun Kwak. 3d human pose estimation using convolutional neural networks with 2d pose information. In *Eur. Conf. Comput. Vis.*, pages 156–169. Springer, 2016.

Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7025–7034, 2017.

Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 459–468, 2018.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10975–10985, 2019.

Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Trans. Graph.*, 40(4), 2021.

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.

Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A Kakadiaris. 3d human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding*, 152:1–20, 2016.

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 *Eur. Conf. Comput. Vis.*, pages 461–478. Springer, 2022.

Mingyi Shi, Kfir Aberman, Andreas Aristidou, Taku Komura, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. *arXiv preprint arXiv:2006.12075*, 2020.

Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1-2):4–27, 2010.

Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation via actor-critic gpt with choreographic memory. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Int. Conf. Comput. Vis.*, pages 2602–2611, 2017.

Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Eur. Conf. Comput. Vis.*, pages 529–545, 2018.

Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Int. Conf. Comput. Vis.*, pages 11179–11188, 2021.

Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13243–13252, 2022.

Zhenhua Tang, Zhaofan Qiu, Yanbin Hao, Richang Hong, and Ting Yao. 3d human pose estimation with spatio-temporal criss-cross attention. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4790–4799, 2023.

Hongyu Tao, Shuaiying Hou, Changqing Zou, Hujun Bao, and Weiwei Xu. Neural motion graph. In *SIGGRAPH Asia 2023 Conference Papers*, New York, NY, USA, 2023. Association for Computing Machinery.

Bugra Tekin, Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *Int. Conf. Comput. Vis.*, pages 3941–3950, 2017.

Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2500–2509, 2017.

Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Int. Soc. for Mus. Inform. Retr. Conf.*, pages 501–510, Delft, Netherlands, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Adv. Neural Inform. Process. Syst.*, 30, 2017.

Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7782–7791, 2019.

Jue Wang, Shaoli Huang, Xinchao Wang, and Dacheng Tao. Not all parts are created equal: 3d pose estimation by modeling bi-directional dependencies of body parts. In *Int. Conf. Comput. Vis.*, pages 7771–7780, 2019.

Jinbao Wang, Shujie Tan, Xiantong Zhen, Shuo Xu, Feng Zheng, Zhenyu He, and Ling Shao. Deep 3d human pose estimation: A review. *Computer Vision and Image Understanding*, 210:103225, 2021.

Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5255–5264, 2018.

Yu Zhan, Fenghai Li, Renliang Weng, and Wongun Choi. Ray3d: Ray-based 3d human pose estimation for monocular absolute 3d localization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13116–13125, 2022.

Jianfeng Zhang, Dongdong Yu, Jun Hao Liew, Xuecheng Nie, and Jiashi Feng. Body meshes as points. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 546–556, 2021.

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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13232–13242, 2022.

Juze Zhang, Ye Shi, Yuexin Ma, Lan Xu, Jingyi Yu, and Jingya Wang. Ikol: Inverse kinematics optimization layer for 3d human pose and shape estimation via gauss-newton differentiation. In *AAAI*, pages 3454–3462, 2023a.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023b.

Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3425–3435, 2019.

Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8877–8886, 2023.

Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *Int. Conf. Comput. Vis.*, pages 11656–11665, 2021.

Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM Computing Surveys*, 56(1):1–37, 2023.

Kun Zhou, Xiaoguang Han, Nianjuan Jiang, Kui Jia, and Jiangbo Lu. Hemlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation. In *Int. Conf. Comput. Vis.*, pages 2344–2353, 2019a.

Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei. Deep kinematic pose regression. In *Eur. Conf. Comput. Vis.*, pages 186–201. Springer, 2016.

Yi Zhou, Connelly Barnes, Lu Jingwan, Yang Jimei, and Li Hao. On the continuity of rotation representations in neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019b.