Towards Robust and Effective Human Pose Estimation and Generation

Zhongyu Jiang

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:
Jenq-Neng Hwang, Chair
Linda Shapiro
Rania Hussein

Program Authorized to Offer Degree:
Electrical and Computer Engineering

University of Washington

Abstract

Towards Robust and Effective Human Pose Estimation and Generation

Zhongyu Jiang

Chair of the Supervisory Committee:
Jenq-Neng Hwang
Electrical and Computer Engineering

Human pose estimation (HPE) in both 2D and 3D remains a fundamental yet challenging problem in computer vision, with broad applications in action recognition, human-computer interaction, motion analysis, and object tracking. Despite recent advances, achieving robustness and efficiency in real-world and edge-device scenarios remains difficult. This dissertation presents a series of contributions toward making HPE more effective and robust. Specifically, we propose (1) a temporal-based 2D HPE method for golf swing analysis, (2) an optimization-driven pipeline for 3D HPE, and (3) a unified contrastive learning-based framework for 2D-3D pose representation. Furthermore, building upon HPE, we explore its potential in human motion generation. In particular, we introduce PackDiT, a novel diffusion-based framework for joint motion and text generation via mutual prompting. PackDiT effectively integrates text and motion generation by leveraging a unique training strategy with two DiT models (Text-DiT and Motion-DiT) with shared latent spaces, enabling text-to-motion, motion-to-text, and joint motion-text synthesis. Evaluated on the HumanML3D dataset, PackDiT outperforms state-of-the-art generative models across multiple tasks, demonstrating its capability as a unified framework for motion understanding and generation. The dissertation discusses challenges, limitations, and potential directions for advancing HPE and human motion generation in future research.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Prof. Jenq-Neng Hwang, for his invaluable guidance, support, and mentorship throughout my research journey. His expertise, encouragement, and insightful feedback have been instrumental in shaping my work and academic growth.

I am also sincerely grateful to my committee members, Prof. Linda Shapiro, Prof. Rania Hussein, and Prof. Chengcheng Zhu, for their time, thoughtful advice, and constructive feedback. Their expertise and perspectives have greatly enriched my research and helped me refine my work.

Finally, I would like to thank my colleagues, friends, and family for their constant support and encouragement throughout this journey.

DEDICATION

to my Mom, Dad and my dear wife, Xingyun.

Chapter 1

INTRODUCTION

Human Pose Estimation (HPE) constitutes a fundamental task in computer vision with widespread applications, including multiple object tracking[2, 147], action recognition[144, 32], human-computer interaction[189, 60], human body reconstruction[192, 84], and sports analytics [217, 64]. 2D HPE involves estimating human keypoints from image data and is generally categorized into top-down and bottom-up approaches based on their architectural design. In contrast, 3D HPE methods can be broadly classified into 2D-to-3D lifting and image-to-3D regression frameworks.

Despite recent progress, deploying HPE in real-world applications remains challenging due to domain shifts, high computational costs, and robustness issues. This dissertation introduces several contributions aimed at improving the efficiency, robustness, and generalization of 2D and 3D HPE, followed by an exploration of motion generation as a downstream task.

In Chapter 3, we propose a lightweight temporal-based 2D HPE pipeline designed for golf swing analysis, **GolfPose**. Accurate pose estimation is crucial for understanding and evaluating the motions of golf players. However, 2D HPE in golf presents unique challenges, including motion blur and severe self-occlusion. By integrating temporal information within a computationally efficient pipeline, we develop an accurate, lightweight model capable of running on edge devices, making real-time golf swing analysis more accessible.

In Chapter 4, we introduce **ZeDO**, a novel optimization-based 3D HPE pipeline that eliminates domain gaps between training and testing data inherent in learning-based approaches. By leveraging a pre-trained diffusion-based pose generation model, ZeDO iteratively refines 3D poses through minimization of the 2D re-projection error, achieving state-of-the-art performance compared to fully supervised learning-based methods.

After exploring various 2D and 3D pose estimation methodologies, an important question

arises: can human pose representations be unified across different modalities? Existing methods often treat 2D and 3D HPE as distinct tasks, limiting cross-modal generalization and requiring separate models for each domain. Unifying these representations could enable more efficient and transferable HPE models, improving both performance and robustness across diverse settings.

In Chapter 5, we present **UniHPR**, a unified human pose representation learning framework that bridges the gap between 2D and 3D human pose estimation. Existing HPE methods often treat 2D and 3D pose estimation as separate tasks, leading to inefficiencies and a lack of cross-modal generalization. To address this, UniHPR leverages contrastive learning to align embeddings from 2D poses, 3D poses, and images within a shared feature space, enabling a single unified pipeline for both 2D and 3D HPE. By incorporating singular value-based contrastive learning, UniHPR enhances feature alignment and improves robustness across diverse datasets and domains, demonstrating state-of-the-art performance in multi-modal HPE tasks.

With the advancements in pose representation learning, a natural progression is to explore how these representations can be leveraged for more representations, e.g., text. By modeling the relationships between human poses and texts over time, it is possible to generate plausible motion trajectories and synthesize movements. Moreover, the synthesized motions can serve as augmented data to enhance the performance of HPE methods.

Chapter 6 explores motion generation and language-guided motion synthesis. We introduce **PackDiT**, a diffusion-based framework for joint motion and text generation via mutual prompting. Unlike existing methods that perform unidirectional translation (i.e., text-to-motion or motion-to-text in isolation), PackDiT employs two Diffusion Transformers (DiTs) that interact via a shared latent space, enabling bidirectional generation and multi-modal synthesis. Through experiments on HumanML3D, we demonstrate that mutual prompting significantly improves coherence and diversity in motion-text alignment, setting a new benchmark for multi-modal motion understanding and generation.

Chapter 2

RELATED WORKS

2.1 2D Human Pose Estimation

Deep learning has proved its superior performance in many vision tasks, including pose estimation. Initially, the positions of keypoints can be regressed directly from the cropped human images [175], while later on, estimating keypoint heatmaps [191, 120] followed by choosing the locations with the highest values as the keypoint coordinates becomes mainstream methods with better accuracy, called Top-Down methods. To accelerate the 2D HPE, openpose [13, 14] proposes Bottom-Up based 2D HPE by estimating all keypoints across the whole images without cropping and applying the association of estimated keypoints later.

2.1.1 Top-Down Methods

Top-down methods[191, 120, 24, 21, 194, 159, 187] first detect bounding boxes of humans and then estimate keypoints for each detected bounding box, as shown in Fig 2.1(Adopted from [19]). 2D HPE relies on both local and global understanding of images, and therefore, researchers [191, 120, 24] focus on increasing the receptive field of the convolution neural networks (CNN) by using large convolution kernels and deepening networks with skip connections and residual modules [54]. Then, inspired by Feature Pyramid Network [94], Cascade Pyramid Network (CPN) [21] utilizes features from different dimensions and receptive fields together and better estimates 2D human poses. Following CPN, HRNet [159, 187] further extends FPN [94] and Stacked Hourglass [120] and achieves state-of-the-art performance. Instead of merging features from different dimensions gradually, like Hourglass networks [120], or at the end of networks [21], HRNet gradually generates features in lower dimensions and persists features in different dimensions in parallel. Therefore, the intermediate features in HRNet are fused with different dimensions stage by stage

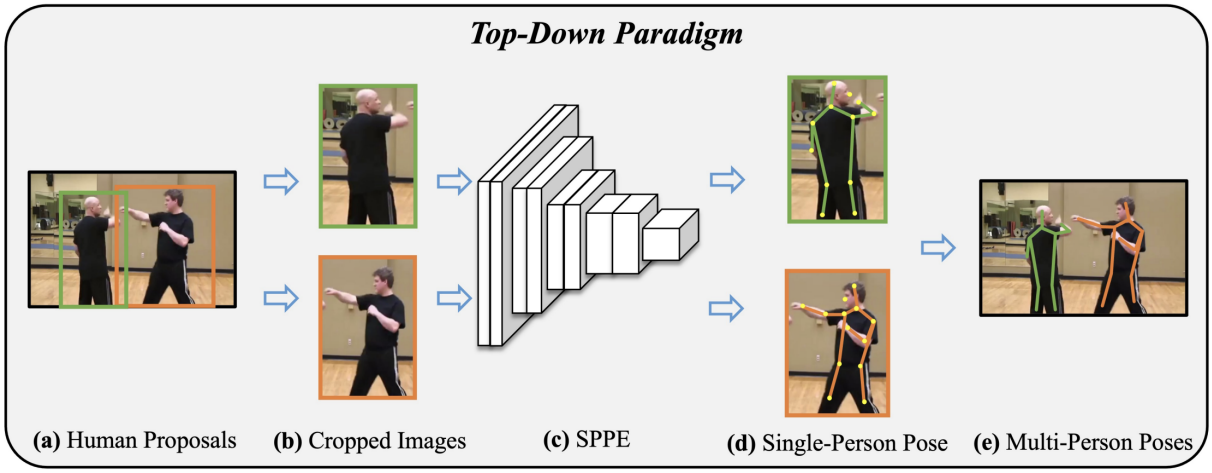


Figure 2.1: The paradigm of Top-Down methods. First, the human proposals are detected and cropped. Then, cropped images are sent to the Single Person Pose Estimation (SPPE) network to estimate poses.

and lead to better 2D HPE performance.

Recently, transformer-based methods show their impressive performance on computer vision tasks. ViTPose [197] adopts the architecture of Vision Transformer [31], but integrates the patch features instead of the [CLS] token to estimate 2D human poses via predicting heatmaps. Furthermore, HRFormer [201] is able to integrate the spirit of HRNet into transformer architectures and achieves comparable results with HRNet using less amount of parameters and computation.

2.1.2 Bottom-Up Methods

Although Top-Down methods achieve numerous successes in 2D HPE, they estimate human pose instance by instance, which significantly impacts the inference speed in real-world applications. Openpose [14, 13] firstly proposes the Bottom-Up approach to estimate 2D poses of all humans in a frame at once. Bottom-Up methods [14, 119, 124, 13, 224, 22, 34] estimate all keypoints at once and conduct keypoint association to group keypoints of each person. Openpose [14, 13] estimates the keypoint heatmaps and part affinity fields, which represent the connection between adjacent

joints. Different from Openpose, HigherHRNet [22] introduces person tags for the keypoint association. With the help of a powerful backbone, HigherHRNet achieves better performance. CenterNet [224] adopts a different strategy. Instead of keypoint estimation and association, CenterNet detects humans as points and simultaneously regresses the keypoint offsets from the center of each human.

2.2 3D Human Pose Estimation

2.2.1 Optimization-based Methods

Optimization-based methods, which estimate 3D poses frame-by-frame and case-by-case, are not handicapped by domain gaps or varying camera intrinsic parameters, but their performance so far is much worse than learning-based methods. Utilizing the SMPL[101] model, SMPLify[7] is capable of optimizing the 3D human poses by minimizing the 2D keypoint re-projection error and satisfying lots of kinematic constraints. Furthermore, Müller *et al*[118] propose SMPLify-XMC as an improved version of SMPLify with more constraints about the human body and more inputs including height and age. Zheng *et al*[168] propose an optimization-based hierarchical 3D human pose estimation pipeline that can estimate both 3D human pose and locations at the same time. Recently, Song *et al*[153] and Choutas *et al*[23] focus on train an optimizer to fit the SMPL model to estimated 2D human poses.

2.2.2 Lifting-based Methods

2D-3D lifting networks employ either a single frame or a sequence of normalized 2D keypoints as input to generate corresponding 3D keypoints[112, 129, 199, 215]. Pavllo *et al*[129] leverage dilated temporal convolutions with semi-supervised ways to improve 3D pose estimation in videos. Yang *et al*[199] enhance 3D human pose estimation by leveraging in-the-wild 2D annotations and a novel refinement network module in a weakly-supervised framework. Li *et al*[86] propose a scalable data augmentation technique that synthesizes unseen 3D human skeletons for training 2D-to-3D networks, effectively reducing dataset bias and improving model generalization to rare

poses. These methods, despite the need of two-stage processing to obtain the 2D keypoints in advance, have demonstrated superior performance on several benchmark datasets and are highly efficient, especially when adapted for temporal considerations.

2.2.3 Image-based Methods

End-to-end 3D HPE methods directly transform image data into 3D pose representations, such as those put forth by Guler *et al*[48], Tung *et al*[177], Tan *et al*[167], and various other research teams including those behind SPIN[76], ROMP[160], BEV[162], and CLIFF[91], who successfully utilize scale and variable height information, effectively resolving issues of depth/height ambiguity. For instance, the methodology introduced by Sun *et al*[160] is a one-stage process that allows for real-time, monocular 3D mesh recovery of multiple individuals. Further contributing to this field, Sun *et al*[162] develop a single-shot method capable of simultaneously regressing the pose, shape, and relative depth of multiple people within a single image, utilizing the Bird's-Eye-View representation for depth reasoning while accommodating variable heights. Within the area of 3D pose estimation from single images, these one-stage techniques consistently demonstrate robust performance despite their comparatively streamlined architectural designs.

2.3 Transformer

Transformer [180] is originally designed to solve sequential input of variable length, especially for Natural Language Processing (NLP) tasks. However, after several years, transformers are proven to achieve superior performance on Computer Vision tasks as well, such as Image Classification [31, 99], Object Detection [15, 227], Multi-Object Tracking [10, 115], and Human Pose Estimation [197, 41, 225, 218]. Taking advantage of self-attention mechanism, the transformer is able to find the connection between all input tokens, which can better generate global features compared with convolution kernels. Furthermore, for HPE tasks, researchers utilize the transformer on spatial and temporal information separately, which improves the performance and minimizes the computational cost.

2.4 Diffusion Model

For many years, researchers have been eager to find an effective method to generate various kinds of data, *e.g*, text, images, audio, and *etc* After the creations of VAE [73], GAN [45], Normalizing Flows [138] and *etc*, diffusion models [55, 150, 154, 157, 16, 4, 12, 11] are proposed and shown to provide the best quality of generated results by training the model to gradually denoise the randomly initialized noise data and generate the final result. Based on diffusion probabilistic model (DPM) [148], Ho *et al*proposed denoising DPM (DDPM), which utilizes U-Net [141] to denoise the noisy data step-by-step to recover the original data. During forward diffusion of a DDPM, noisy data are generated by adding Gaussian noise to the original data step-by-step. In contrast, reverse diffusion aims to predict and remove the added Gaussian noise and gradually recover the original data. DDIM [150] is then proposed to accelerate the reverse diffusion of DDPM by skipping certain steps, and Score Matching Network [157] takes advantage of Stochastic Differential Equations (SDE) to build a more general and effective diffusion pipeline. To further scale up the diffusion model, Peebles *et al*[130] propose the Diffusion Transformer (DiT), which utilizes transformers as the backbone of diffusion models.

2.5 Multi-Modal Learning and Feature Alignment

How to align data from multiple modalities is a challenging and important task. CLIP [134] is trained on web-scale text-image pairs under a contrastive learning paradigm. Inspired by the outstanding capability of learning representations for both vision and language of CLIP, various models have adopted the contrastive method to pursue zero-shot performance in other areas, such as [40, 53, 97, 106], by pre-training the model which maximizes cross-modal similarity scores. In this process, CLIP-based models would automatically learn implicit multi-modal alignments, which intensively reduces the difficulty of manually building feature correspondence. However, the majority of prior research has predominantly concentrated on visual-language or other visual-related cross-modal capacity, primarily due to the substantial availability of image-text paired datasets. Nonetheless, few works focus on the area of human pose. MPM [214] aims to learn

shared 2D-3D human pose features by the masked modeling paradigm. Yet, there are limited instances where the correlation between 2D/3D human pose has been clearly researched using a contrastive paradigm similar to our method.

2.6 Text-condition Human Motion Generation

Human motion generation aims to generate realistic and controllable human pose sequences. Usually, people adopt SMPL [102] or keypoint [17, 66, 98] as the representation of 3D human pose instead of 3D keypoint joints. Researchers have been exploring using text, action, audio, music, or even scenes and objects as the conditions to guide the human motion generation. Among all those conditions, text has a remarkable capacity to convey information related to various actions, speeds, directions, and destinations, either explicitly or implicitly. This feature makes the text an appealing medium for generating human motion. Text2Action [1] is the first to leverage GAN to generate a variety of motions from a given natural language description. Recently, diffusion models have been adopted to motion generation tasks successfully as well [28, 49, 59, 211, 212]. However, although these methods have achieved excellent results in motion generation, they cannot simultaneously accomplish the task of action understanding, such as motion-to-text. Recently, MotionGPT [63] uses the autoregressive paradigm of transformers to unify text-to-motion and motion-to-text within a single framework.

Chapter 3

GOLFPOSE: GOLF SWING ANALYSES WITH A MONOCULAR CAMERA BASED HUMAN POSE ESTIMATION

3.1 Introduction

Sports are social-cultural activities and have already become important parts of our daily life. It allows people to interact with each other regardless of their social status, helps to improve the quality of people's lives, and also serves as a significant symbol for measuring the development and progress of a country and society.

A significant amount of resources have been allocated to the modern sports industry, which demands higher requirements not only on the athletes themselves but also in a lot of related supporting technologies. For example, tracking players' trajectories in the field[8] can improve the audiences' experience during game broadcasting, analyze and assess players' performance for better coaching[139], detect and prevent life-threatening situations to players, etc. These requirements call for accurate analyses of actions, conditions, and environments across different players, scenarios, and sports events.

Before the modern sports industry era, sports analytics could only use naked human eyes and their own experience to measure and analyze, which are unreliable, inefficient, and too subjective to generalize to different players, scenarios, and sports events. Therefore, manual analyses of sports are being gradually replaced by a combination of different sensors and algorithms that can automatically do all the cumbersome analyses. These capacities can help better assess the crucial sports event moments, resulting in more precise, efficient, and generalizable analysis results.

Among these newly emerging technologies, rapid development in computer vision communities combined with recent deep learning technologies have been highly appreciated in terms of efficiency and accuracy. Furthermore, thanks to the popularity of social media and online stream-

Figure 3.1: Two sample images from our dataset. The right one is annotated with 38 keypoints.

ing, massive video and image data are generated and become available for researchers to utilize and improve the performance of their applications.

More than 24.8 million people played golf in the U.S. in 2020, and the demand for user-friendly automated golf swing analyses is rising. The crucial thing for sports analyses is how to understand and judge the motion of sports players. Therefore, an accurate and efficient human pose estimation (HPE) method is critical for reliable golf swing analyses. However, HPE for golf swing analyses is different from the other HPE tasks, because of the input format, motion blur and self-occlusion. Therefore, in this paper, we propose a temporal-based lightweight 2D HPE pipeline, called GolfPose, which can be running on mobile devices for golf swing analyses.

Our contributions include:

- A light-weight monocular temporal-based 2D human pose estimation model which provides

accurate pose estimation results and can be deployed on mobile devices.

- Incorporating line segment based golf club detection(GCD) to further improve pose estimation accuracy.
- An annotated golf swing dataset with more than 500 videos of over 120 fps and 120,000 images.

3.2 Method

GolfPose is targeted at golf-playing scenarios. Its goal is to generate accurate 3D pose estimation from a monocular swing video taken from mobile devices. The 2D GolfPose first generates reliable keypoints on both players' body and golf club, which are then systematically converted to 3D poses for further analyses. In this paper, due to the page limitation, we mainly focus on the innovations and performance improvements made to the 2D architecture of GolfPose. Our model is constructed to address this problem: First, we build a CNN-based temporal 2D HPE model based on an existing image-based HPE framework. Since our input is a short clip of a video sequence instead of a single image, we are able to utilize temporal information to increase the accuracy of keypoint prediction. Then, we implement a line segment algorithm, a traditional computer vision technique, to fix inaccurate predictions on golf club keypoints generated from the 2D HPE model. The overall pipeline of our 2D GolfPose architecture is illustrated in Figure 3.2. In the following sections, we will introduce the main components of our 2D GolfPose in detail.

3.2.1 Problem Formulation

Let $S \in \mathbb{R}^{L \times H \times W \times C}$ be the input video of L RGB frames ($C = 3$) with $H \times W$ size. Our goal of HPE is to predict a set of 2D keypoints $J \in \mathbb{R}^{N \times 2}$ for every frame in the video sequence, where N denotes the number of keypoints, which is $N = 38$ in our dataset. Our approach is sequence-based, i.e., in every time step t, it operates on a short video clip: $V = \{F_{t-n}, ..., F_t, ..., F_{t+n}\}$ and outputs the HPE result for the center frame F_t.

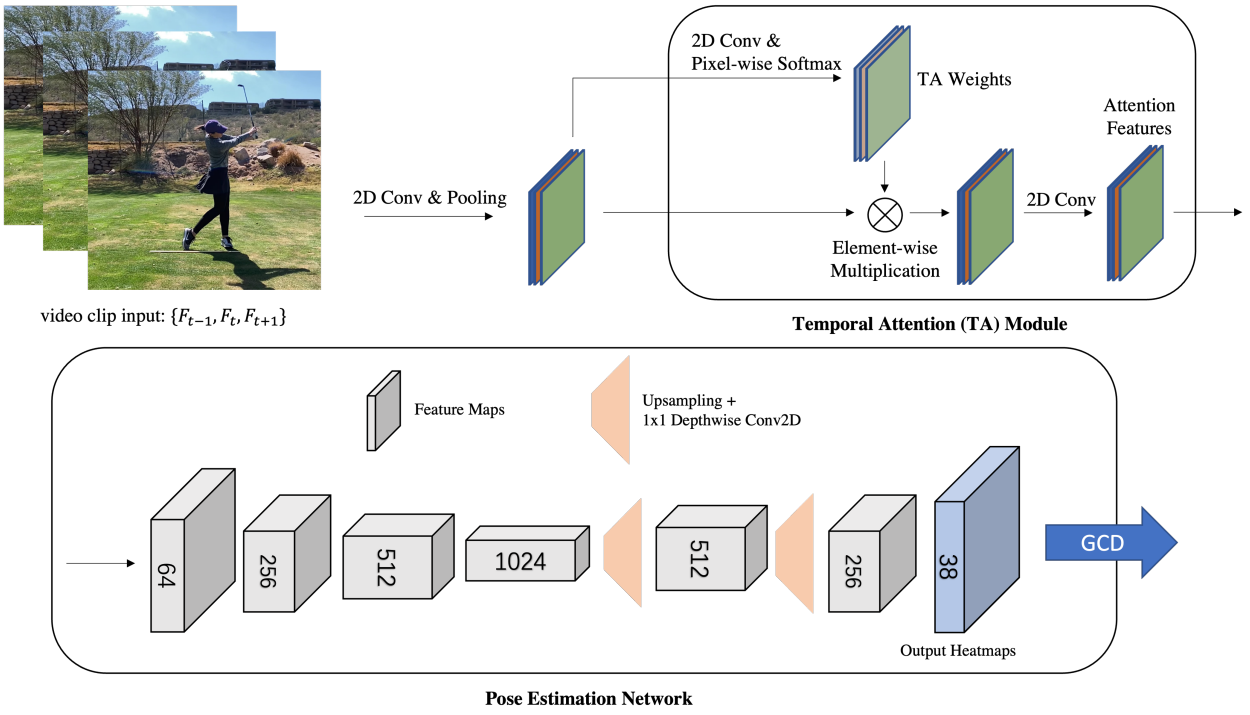


Figure 3.2: Overall pipeline of our proposed GolfPose for human pose estimation. Three frames are used as the input, for example. First, there is a lightweight CNN-based 2D HPE model with temporal attention to utilize temporal information and increase the accuracy of occluded or fast-moving keypoint estimation. Then, golf club detection (GCD) is applied to fix inaccurate predictions on golf club keypoints generated from the 2D HPE model.

3.2.2 2D Temporal based Model

To address self-occlusion and motion blur problems, temporal information is a good solution since the occluded keypoints may be visible in adjacent frames, and the blurry keypoints can be recovered based on multiple consecutive frames. Therefore, we propose a 2D temporal-based HPE model to generate accurate keypoints.

The temporal attention module is modified from [9], where the original temporal attention module is based on 3D convolutions, which incur high computational costs and cannot be performed real-time on mobile devices. In order to solve the latency issue, we replace the 3D convolutions

with depthwise 2D convolutions and modify the architecture as well. As shown in Tables 3.1 and 3.4, this modification improves the inference speed significantly on mobile devices with only a small amount of accuracy degradation. The rest of the network is modified from LPN[213] by replacing the mobile incompatible operations.

Algorithm 1 Golf Club Detection

Require: Image I_t, Position of club mid hands J_{md}, top of handle J_{handle}, hosel J_{hosel}

Ensure: Updated hosel position J_{hosel}

$\vec{D}_{init} \leftarrow J_{md} - J_{handle}$

Compute bounding box covering whole golf club and crop out this region

Apply Line Segment Detection from OpenCV and assign results as list of line segments, L

for all $line$ in L **do**

  if $\langle line, \vec{D}_{init} \rangle >= 15°$ **then**

    remove $line$

  end if

end for

Find N line segments with the smallest distance to J_{md}

Set searching space as all line segments that are not included in any of these N line segments

for all $line$ in N line segments **do**

  while searching space $\neq \varnothing$ **do**

    $nline \leftarrow$ line segment with minimum distance $mdist$ to $line$

    if $mdist < thre$ **then**

      $line \leftarrow$ connect start of $line$ with end of $nline$.

      remove $nline$ in searching space

    else

      break

    end if

  end while

end for

$L_{club} \leftarrow$ longest $line$ from N lines.

if $\langle L_{club}, \vec{D}_{init} \rangle < 15°$ **then**

  $J_{hosel} \leftarrow$ projection of J_{hosel} in the direction of L_{club}

else

  $J_{hosel} \leftarrow$ end point of L_{club}

end if

To train the model, like other HPE model, we adopt the mean square error loss to minimize the

difference between the predicted keypoint heatmaps H_{pred} and the ground truth keypoint heatmaps H_{GT}:

$$L_H = \|H_{pred} - H_{GT}\|_2^2 \tag{3.1}$$

3.2.3 Golf Club Detection

Fast movement of the golf club during the swing can cause issues in the 2D GolfPose model, such as missing detection of golf club due to motion blur. Moreover, golf club can move out of image boundary, which may also cause detection failure. These issues greatly hinder the performance of keypoints prediction, especially the points on the club hosel, which is crucial for analyzing golf swing. To address the problem of inaccurate hosel prediction, we resort to traditional computer vision techniques.

Algorithm 1 describes the proposed golf club detection (GCD) algorithm. Based on the prediction results from the 2D GolfPose keypoints model, we first determine the bounding box of the golf club and then apply the line segment detection (LSD) algorithm [82] to the cropped golf club region. The output of the LSD gives us lots of unconnected and short line segments pointing in various directions since LSD only depends on pixel information and can be easily influenced by other line-shape elements in the environment, like grass and ground.

We set the direction from the club top of the handle J_{handle} to the club middle hand J_{md} as the reference direction. With previously detected redundant and erroneous line segments and the reference direction, we can first remove those line segments which have large angles with respect to the reference direction and assign the vector directions of the remaining segments.

Next, we implement an iterative process to remove outliers that do not lie on the golf club and connect different vectors to form the whole golf club line. In every iteration step, we find a new vector in the search space to connect to current club line candidates and form a new line representing the golf club. The connected vector should have a consistent direction with the current club line with the minimum distance between these two vectors. The distance D between two vectors is the distance between the first one's end point and the second one's starting point.

3.3 Experimental Results

In this section, we evaluate the performance of our proposed GolfPose system. We mainly use the golf swing dataset we collected ourselves for benchmarking. Details of the dataset and training process as well as system performance will be discussed below:

3.3.1 Dataset and evaluation metrics

As far as we know, we are the first to apply deep learning based vision techniques to swing analyses in golf playing scenarios. In order to get reliable pose estimation results for golf swings, we use a subset of collected and annotated dataset for our performance evaluation. This subset dataset includes 120,000 images, where 100,000 images are used for training and 20,000 images for validation and testing which are not used in the training. Unlike traditional 2D human pose estimation datasets, e.g., COCO[95], our dataset is video based and is recorded with over 120fps to eliminate the motion blur. There are in total 38 keypoints annotated in our dataset, as shown in the right image of Figure 3.1, including some keypoints on the golf club. Since video resolution and the size of recorded players in our dataset are relatively similar, we use 2D mean pixel error (MPE) as our evaluation metrics to evaluate the accuracy of the keypoint localization.

3.3.2 Training details

Since our GolfPose is designed for mobile devices, we implement the system using the publicly available TensorFlow framework and convert it to TensorFlow Lite model for mobile inference.

The pose estimation model is trained in an end-to-end manner. All parameters are initialized randomly from the zero-mean Gaussian distribution with $\sigma = 0.001$. We use Adam optimizer with a mini-batch size of 32 to update the parameters. The total number of training epochs is 150, and the initial learning rate is set to 0.001, reduced by a factor of 10 at the 90^{th} and 120^{th} epoch.

The detected and cropped human bounding boxes from the video frames are fixed to a certain aspect ratio (i.e., height: width = 4:3). The cropped bounding box is resized to 256×192, keeping the original aspect ratio and padding with the black background, and served as the input image. In

Method	Input size	#Params↓	GFLOPs↓	MPE^A ↓	MPE^B ↓	MPE^C ↓
HRNet-W32[159]	256 × 192	28.5M	7.1	9.20	**8.48**	13.98
LPN[213]	256 × 192	2.9M	2.28	9.60	9.34	11.34
Ours (Conv3D)	256 × 192 × 3	2.8M	8.22	**9.08**	9.08	9.10
Ours (Conv2D)	256 × 192 × 3	3.2M	5.46	9.15	9.18	**8.95**
Ours (Depthwise Conv2D)	256 × 192 × 3	2.9M	3.42	9.16	9.15	9.21

Table 3.1: Experimental results on our golf swing test set. MPE^A stands for MPE of all keypoints. MPE^B for MPE of body keypoints. MPE^C for MPE of club keypoints

addition to common data augmentation operations like random rotation, random scale, and flipping, we also add several additional augmentation operations, aiming at increasing the system robustness under specific conditions. For example, we randomly adjust the input image brightness to represent overexposed or underexposed environments. We also add random Gaussian noise and multi-frame averaging to the original image to simulate motion blur caused by camera shaking. Our training are performed on one NVIDIA 1080Ti GPU, and the training takes about 36 hours to complete.

3.3.3 Results

As shown in Table 3.1, we test our system performance on the 20,000 images in the validation and testing split. Compared to the HRNet[159] with the same input resolution, our GolfPose model can achieve better performance with much less number of parameters and fewer GFLOPs. Furthermore, for keypoints on the golf club, with the help of multi-frame temporal input and our temporal attention module, our performance is much better than the image-based state-of-the-art method, HRNet. Figure 3.4 shows some representative qualitative performance of the keypoints estimated by the proposed 2D GolfPose. In addition, our model is also favorable in terms of inference speed when deployed in mobile devices because of the smaller model size, and all the operations in the architecture can be accelerated by mobile GPUs.

Golf Club Detection (GCD) is incorporated to correct the failing cases from the HPE model

and to improve golf club keypoint estimation accuracy. According to Table 3.2, GCD significantly decreases the golf club hosel's standard deviation of Pixel Error (Std. PE), from 33.89 to 23.19, which means GCD does correct the wrong estimation of the model output for failure cases.

Method	$\text{MPE}^{hosel} \downarrow$	$\text{Std. PE}^{hosel} \downarrow$
Model	11.93	33.89
GCD	**10.76**	**22.39**

Table 3.2: Performance improvement after GCD.

3.3.4 Ablation Study

We study the effect of each component in our methods, including the length of sequence input and different designs of the temporal attention module. All results are evaluated on our collected validation and testing set and with the same input size (256 x 192) and same training scheme.

Length of input sequence: Since our model is sequence-based instead of single image-based, we would like to explore the how input sequence length can affect the accuracy and inference speed of the model and eventually obtain a good trade-off between these two metrics. Table 3.3 shows the evaluation results in terms of accuracy and inference speed over input sequence lengths of 3, 5, 7 respectively. We can see that simply increasing the length is not a good choice, it can hurt both accuracy and inference speed.

Length	$\text{MPE}^{A} \downarrow$	$\text{Std. PE}^{A} \downarrow$	Inference Speed
3-frame	9.159	8.922	**50.3ms**
5-frame	**9.056**	**8.560**	50.7ms
7-frame	10.054	9.801	51.6ms

Table 3.3: Model performance with different input sequence length. Test on Samsung S20 Ultra.

Temporal attention module: In order to furthermore improve the inference speed of the

temporal-based model, we also modify the original temporal attention module by replacing the 3D convolution with 2D convolution or depthwise 2D convolution. According to Table 3.1 and Table 3.4, the GFLOPs and inference time is significantly decreased with the modification while maintaining the similar performance, which shows that depthwise 2D convolution based temporal attention module is the best choice for mobile device inference, with a good balance between accuracy and inference speed.

Method	Inference Speed	GFLOPs
Single Frame	**27.6ms**	2.28
Conv3D	130.0ms	8.22
Conv2d	68.5ms	5.46
Depthwise Conv2D	**50.3ms**	3.42

Table 3.4: Inference speed per frame of different temporal-based models and the image-based model. Test on Samsung S20 Ultra.

3.4 Conclusion

In this paper, we propose a novel lightweight temporal-based 2D human pose estimation method, GolfPose, and a golf club detection method for further improving keypoint prediction accuracy on the golf club, which can be deployed on mobile devices for efficient and accurate golf swing analyses. The success of this pipeline is under the assumption that players are not moving,which is justified in golf swinging while not applicable to moving around players in other sports that require the human tracking mechanism to be added for temporal-based human pose estimation. In the future, we will work on migrating our method to other sports with the help of human tracking.

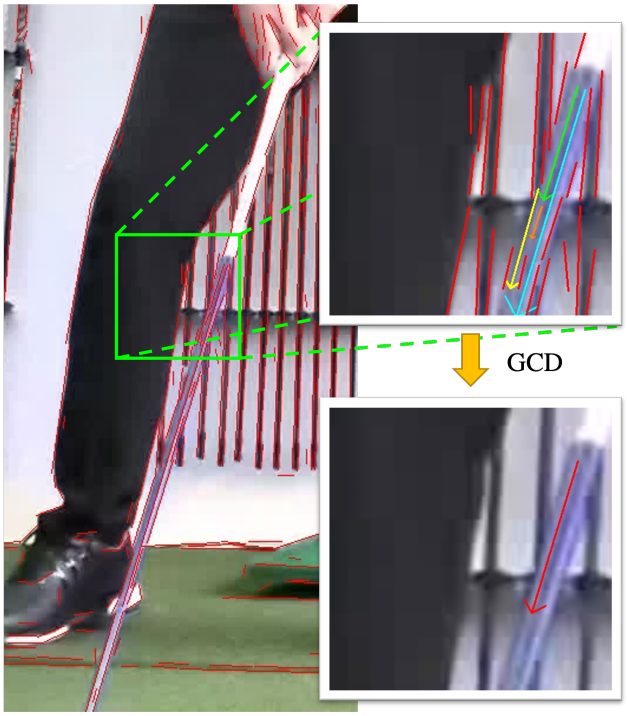


Figure 3.3: Golf club detection (GCD). After the initial line segment detection, there are many noisy line segment detection results. The GCD first filters line segments with the help of J_{md} and J_{handle}, and then finds some potential starting line segments to form the golf club. As the figure shows, if the green segment is the starting segment and the yellow, orange and blue segments are candidates to be added to the search list, according to the distance between those segments, the orange segment is then closest to be connected the green segment, and they are merged to be the next starting segment.

Figure 3.4: Qualitative results on our dataset.

Chapter 4

BACK TO OPTIMIZATION: DIFFUSION-BASED ZERO-SHOT 3D HUMAN POSE ESTIMATION

4.1 Introduction

As people become increasingly interested in Virtual Reality (VR), Augmented Reality (AR), Human-Computer Interaction and Sports Analysis, 3D Human Pose Estimation (HPE) becomes a crucial component for these applications. Compared with multi-view 3D HPE, monocular-based methods are easier to set up and have lower costs, which are more suitable for VR, AR, and mobile devices. Weng *et al*[192], and Peng *et al*[131] utilize 3D human poses with Neural Radiance Fields (NeRF) for 3D Human Reconstruction. Meanwhile, Bridgeman *et al*[8] propose a 3D HPE and tracking pipeline for soccer analysis, and Jiang *et al*[64] take advantage of 2D and 3D HPE to track the motion of golf players.

With the availability of more benchmark datasets, deep learning-based 3D HPE methods have been shown to outperform all traditional methods and dominate the areas. Combining 2D HPE with SMPL [101] model, Bogo *et al*propose SMPLify[7] as an optimization-based 3D HPE pipeline. 2D-3D lifting [129, 215, 25, 199, 219] and diffusion-based 3D HPE[26, 42] networks leverage 3D human poses from the single-frame or multi-frame 2D poses. On the other hand, Image-to-3D networks [76, 74, 162, 91] estimate 3D human poses directly from images without intermediate 2D human poses. However, as mentioned in [43, 39], learning-based 3D HPE methods suffer from performance degradation with cross-domain or in-the-wild scenarios. During training, these methods implicitly learn camera intrinsic parameters, domain-based 3D human pose distributions, or image features in a certain domain. Although optimization-based methods can mitigate the impact of domain gaps by estimating 3D poses case-by-case, their performances are not comparable to learning-based methods at this moment.

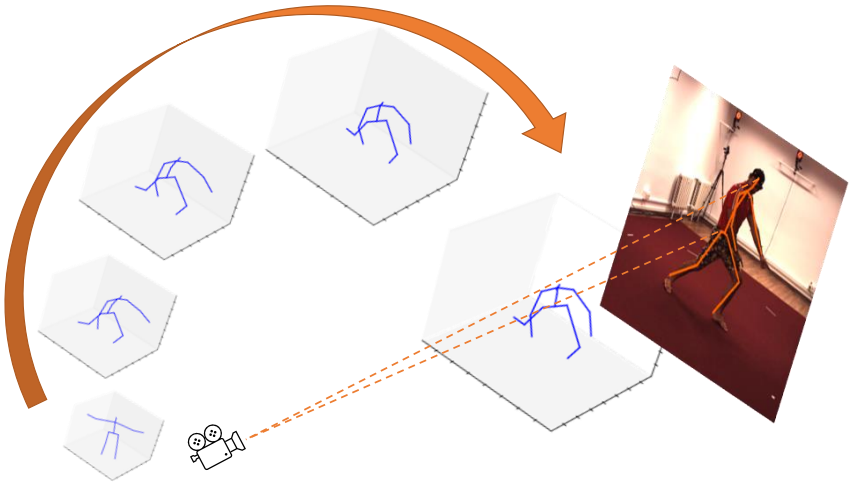


Figure 4.1: ZeDO iteratively estimates 3D poses by minimizing the re-projection error via a diffusion-based method.

To address this problem, Zhan *et al*[204] decouple the camera intrinsic parameters from 2D-3D lifting networks learning by converting 2D keypoints to 2D rays. Gong *et al*[43] and Gholami *et al*[39] generate various 3D poses to bridge domain gaps. Furthermore, Chai *et al*[17] propose a data augmentation pipeline to minimize the 3D human pose spatial distribution gap. However, the above methods still cannot outperform the learning-based 3D HPE. To decouple camera intrinsic and bridge 3D pose domain gaps simultaneously, we propose the Zero-shot Diffusion-based Optimization(ZeDO) pipeline for 3D human pose estimation, which combines a simple yet effective optimization pipeline with a pre-trained diffusion-based 3D human pose generation model.

Different from traditional optimization-based methods[7, 168, 47, 118], which include various kinematic constraints, the diffusion model denoises the output from the optimization pipeline iteratively to ensure optimized poses following human body constraints. Meanwhile, poses are optimized by minimizing 2D keypoint re-projection errors with a simple yet effective optimization pipeline. Thus, ZeDO is able to estimate 3D human poses without training on any 2D-3D or image-3D pairs. Our contributions are as follows:

- The proposed ZeDO pipeline is a Zero-Shot 3D Human Pose Estimation pipeline, which leverages a pre-trained diffusion-based 3D human pose generation model to optimize target 3D poses in the loop during the inference time.

- Compared with other generation and denoising tasks, we take the diffusion model as an optimization tool by combining a simple 3D HPE optimization pipeline with a pre-trained diffusion-based 3D human pose generation model.

- ZeDO achieves state-of-the-art zero-shot 3D HPE performance on Human3.6M, 3DHP, and 3DPW datasets, even on cross-dataset evaluation.

In Sec 4.2, details of our backbone architecture and optimization pipeline are addressed. Experimental results are presented in Sec 4.3, and the ablation studies will be discussed in Sec 4.4. At last, there are conclusions in Sec 4.5.

4.2 Method

As shown in Fig 4.2, ZeDO includes an initial pose optimizer for rotating initial poses and an optimizer in the loop for iteratively optimizing 3D poses.

Firstly, a randomly selected initial 3D pose (a hypothesis), $P_{init} \in R^{J\times 3}$, is rotated to an optimal pose, P_0, by minimizing the re-projection error with detected or ground truth 2D keypoints $p_{2d} \in R^{J\times 2}$. Here, J stands for the number of keypoints. Then, in the ith optimization step, P_i is optimized by the optimizer in the loop, and the pre-trained diffusion model is used to denoise it to P_{i+1} as input of the next iteration. After n iterations, P_n will be the estimated 3D human pose.

Different from other diffusion-based pose estimation methods[26, 42], our diffusion model θ_g is a pose generation model, which is only trained with 3D human poses, and during inference, our diffusion model only takes the optimized pose $\widetilde{P}_i$ and timestamp $t(i)$ as input without any additional pose condition information including 2D poses.

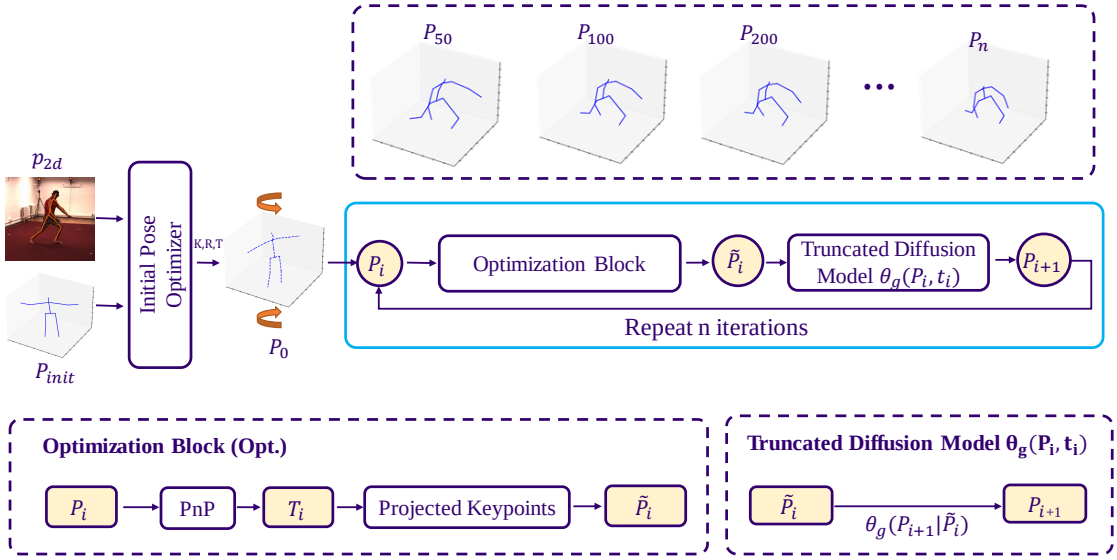


Figure 4.2: The pipeline of ZeDO, which takes an initial 3D pose, called a hypothesis, as input and estimates the pose by minimizing re-projection error with the target 2D pose. After 1000 iterations, ZeDO is able to generate the optimized 3D human poses.

4.2.1 Pre-trained 3D Human Pose Generation Model

We apply the Score Matching [157] on the pre-trained backbone for our 3D human pose generation diffusion model, which rectifies the noisy poses generated after projection to get reasonable 3D poses. During pre-training, the model takes relative-to-pelvis 3D poses $x \in R^{J\times 3}$ as inputs and tries to recover them from recurrent Gaussian noise. In this case, we expect that the diffusion model learns the distribution of real 3D poses and reconstructs $\tilde{x} \in R^{J\times 3}$ to minimize the difference from the inputs. The perturbation strategy used in our Score Matching diffusion model expresses $p(x(t)|x(0))$ in the closed form as:

$$N\left(x(t); x(0)e^{-\frac{1}{2}\int_0^t \beta(s)ds}, \left[1 - e^{-\int_0^t \beta(s)ds}\right]^2 I\right). \tag{1}$$

Besides, built upon the learning strategy of the noise conditional score network (NCSN) [155],

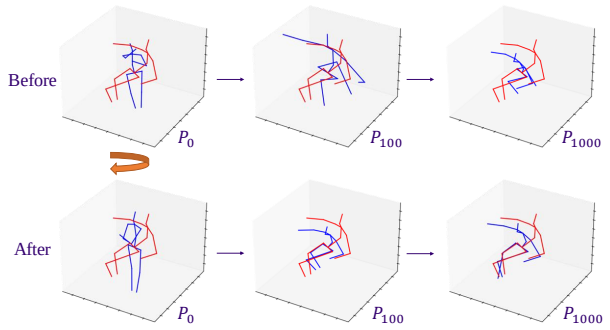


Figure 4.3: The rotation optimization in the initial pose optimizer finds the optimal initial pose and prevents the collapse of the estimated pose.

we formulate our loss function as follows by choosing $\lambda(t) = \sigma(t)^2$:

$$L = E_{U(t;0,1)}\left[\lambda(t)||s_\theta(x(t),t) + \frac{x(t)-\mu}{\sigma^2}||_2^2\right] \tag{2}$$

$$= E_{U(t;0,1)}\left[||\sigma(t)s_\theta(x(t),t) + z||_2^2\right], \tag{3}$$

in which z stands for random noise vector $z \sim N(0,1)$ and s_θ is the pre-trained score matching network. σ represents the variance mentioned in Eq.(1) as $[1 - e^{-\int_0^t \beta(s)ds}]$ and the timestamp or denoising time variable t is uniformly sampled 1000 times from $(0,1]$. The 3D pose generation model is never trained with any 2D-3D or image-3D pairs. We report the results based on the Score Matching based model, similar to GFPose[26]. More results with different backbones, including DDIM[151] and DDPM[56], are shown in Table 4.5.

4.2.2 Initial Pose Optimizer

Similar to other optimization-based 3D HPE methods[7, 168, 47], our optimization pipeline starts from an initial pose, P_{init}. As depicted in Fig 4.3, the optimized pose may not suffice if the initial pose's orientation is significantly different from that of the target pose or even perpendicular to that of the target pose. Therefore, the initial pose optimizer is designed to find the optimal rotation matrix, $R_0 \in SO(3)$, and translation, T_0, of P_{init} by minimizing the re-projection error with 2D

keypoints p_{2d}.

$$\arg\min_{R_0,T_0} \quad \left\| K(R_0 P_{init} + T_0) - p_{2d} \right\|_2 \tag{4.1}$$

$$\text{s.t.} \quad T_{min} \leq T_0 \leq T_{max}, \tag{4.2}$$

where K is the camera intrinsic matrix. After the rotation, the $P_0 = R_o P_{init}$ is the optimal pose sent to the iterative optimization pipeline. As shown in Fig 4.3, the rotation optimization aligns the initial poses with target 2D and 3D poses.

4.2.3 Optimizer in the Loop

In previous works[7, 118, 46, 168], to optimize an accurate 3D human pose, it requires a lot of kinematic constraints, which call for a strong domain knowledge about human motion. Different from those previous works, as shown in Alg 2, our iterative optimizer utilizes the denoising capability and learned human pose prior of the pose generation diffusion model to estimate an accurate 3D human pose with a simple yet effective optimization pipeline without any explicit kinematic constraint.

With a camera intrinsic matrix, K, 2D keypoints, p_{2d}, can be converted to 3D rays, $r \in R^{J\times 3}$, based on perspective projection,

$$r = K^{-1} p_{2d}, \hat{r} = \frac{r}{\|r\|_2}. \tag{4.3}$$

Intuitively, as shown in Fig 4.44.4, projecting the 3D keypoints from P_0 to r will minimize the 2D re-projection error and provide the estimated 3D human poses,

$$\widetilde{P}_0 \leftarrow \Big((P_0 + T) \cdot \hat{r} \Big) \hat{r} - T. \tag{4.4}$$

However, there are two problems: 1) Simply projecting 3D keypoints from P_0 to r generates a noisy 3D human pose, which may not satisfy the kinematic constraints of the human body. 2) With different translations between P_0 and the camera, there are different projection results.

In order to solve these two problems, we need to ensure the estimated 3D poses, P_i, are valid poses and inherently follow the kinematic constraints to find the optimal translation, T_i. This calls for our use of pose prior.

Pose prior Although there is no re-projection error from the optimized poses, the optimized poses may not satisfy the kinematic constraints. Therefore, in previous works, kinematic constraints are added to the pose optimization pipeline, and a complex joint optimization problem is designed to find optimal poses. However, in our pipeline, we take advantage of a pre-trained diffusion-based pose generation model to 'denoise' our optimized poses. As mentioned in DDPM[56], DDIM[151] and Score Matching network[157], the diffusion model is trained by maximizing likelihood, which aims at finding the most possible valid pose based on the input noisy pose. As a result, we use the diffusion-based pose generation model to find the optimal P_{i+1} based on optimized pose $\widetilde{P}_i$,

$$P_{i+1} = \theta_g(\widetilde{P}_i, t(i)), \tag{4.5}$$

which is different from other diffusion-based methods, like GFPose[26] and DiffPose[42], $P = \theta(x, c, t)$, where x is random noise, c is pose condition and t is timestamp.

However, during training, the reverse diffusion starts from the standard Gaussian noise, $\mathcal{N}(0, 1)$, but in our case, the generation model is utilized to denoise an optimized pose, which does not follow the standard Gaussian noise. Inspired by [116, 221], we adopt truncated diffusion model inference, whose timestamp is truncated from the training timestamp during inference. In our case, the timestamp, t, is truncated as $t \in (0, 0.1]$, instead of $(0, 1]$.

Find optimal translations Optimal translations are derived from T_o by minimizing the 2D re-projection error of P_i, depending on the current iteration number, since the optimized pose, P_i, is not reliable enough in the early iterations.

After certain iterations, the optimal translation is derived from the following:

$$T_i \leftarrow \arg\min_{T_i} \left\| C_{2d} \Big(K(P_i + T_i) - p_{2d} \Big) \right\|_2, \tag{4.6}$$

where $C_{2d} \in R^J$ are the confidence scores of 2D keypoints, p_{2d} and K are the camera intrinsic

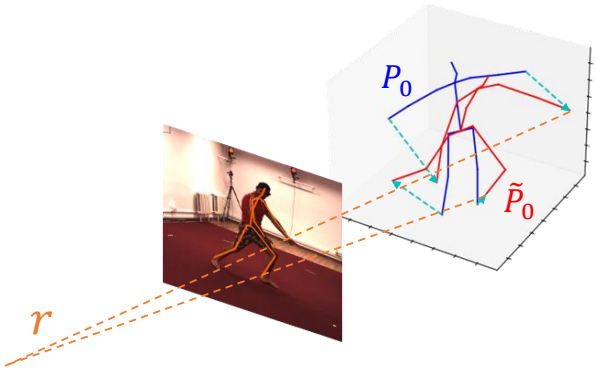


Figure 4.4: By projecting the P_i to r, we minimize the 2D re-projection error and find an optimized pose $\widetilde{P}_i$.

matrices. Inspired by [62], C_{2d} are helpful to guide the translation optimization. There is a closed-form solution to the optimization problem. The details can be found in the supplementary material.

By incorporating optimal translations and kinematic constraints, we can optimize the 3D human poses by iteratively projecting P_i to r and denoising the projected P_{i+1} with the help of the diffusion model.

4.3 Experimental Results

In this section, we will introduce the experimental results of ZeDO on 3DPW[182], Human3.6M[61] and MPI-INF-3DHP[113] datasets. More results on Ski-Pose[139] datasets are included in the supplementary materials. Since ZeDO requires an initial pose as a starting point for optimization, and different initial poses may lead to different pose estimation results, we report our results in Mean Per Joint Position Error (MPJPE) with a single initial pose (single-hypothesis) or minimum MPJPE (minMPJPE) with multiple initial poses, in order to have fair comparison with previous multi-hypothesis 3D HPE methods[26, 190, 145].

4.3.1 Datasets

Human3.6M[61] is the most widely used single-person 3D pose benchmark with more than 3.6 million frames and corresponding 3D human poses. The dataset is collected within a 4 m $\times$ 3 m indoor environment, with 11 professional actors (6 males and 5 females) performing 17 distinct actions such as discussion, smoking, capturing photographs, posing, greeting, and talking on the phone. Following the convention of previous works[129, 215, 86] for fair comparison, we use the S1, 5, 6, 7, and 8 as the training dataset and evaluate the model on S9 and S11.

MPI-INF-3DHP[113] is a large 3D human pose dataset with more than 1.3 million frames captured indoors and outdoors. The 3DHP dataset captures poses of 8 actors, consisting of 4 males and 4 females with 8 different actions each, encompassing a range of activities from simple actions like walking and sitting to more complex exercise poses and dynamic movements. Following [43], we use a sampled 2929 frame test dataset.

3DPW[182] is the first dataset in-the-wild with accurate 3D poses for evaluation. Compared with Human3.6M and MPI-INF-3DHP, 3DPW focuses on outdoor scenarios and captures videos with static and moving cameras. There are 60 video sequences captured in the dataset with 18 different actors. Following [17], we test ZeDO on 3DPW only for cross-dataset evaluation.

4.3.2 Training and Inference Details

We pre-train our 3D pose generation model for 5000 epochs on one NVIDIA A100 with a batch size of 50k, a learning rate of $2e^{-4}$ with an Adam optimizer. The training schedule comes with warmup in the first 5k iterations and cosine learning rate decay in the following iterations. As [157], the timestamp, t, during the forward or reverse diffusion process, is uniformly sampled from $(0, 1]$. All 3D human poses are normalized to pelvis-related coordinates during training and inference. To improve the robustness of the model, we apply flip and rotation data augmentation during training. For cross-domain evaluation, we pre-train the pose generation model in a different dataset and directly test the optimization pipeline without any fine-tuning.

During inference, the pipeline supports single or multiple initial poses. For the initial pose

optimizer, we limit the rotation axis to the z-axis only for better performance. We set the number of warmup iterations as 200, and the number of total iterations as 1000. The timestamp, t, is uniformly sampled from $(0, 0.1]$. The initial poses are sampled from training sets of Human3.6M[61] or 3DHP[113] by the K-Means algorithm. For different numbers of hypotheses, we run K-Means with different numbers of clusters.

4.3.3 Results

Results on 3DPW. 3DPW is a challenging in-the-wild dataset, compared with Human3.6M and MPI-INF-3DHP datasets. 3DPW focuses on outdoor scenarios with both static and moving cameras. For cross-domain evaluation on the 3DPW dataset, we pre-train the pose generation model on the Human3.6M dataset and inference on the 3DPW dataset without any fine-tuning. On 3DPW, we find some of the previous works[85, 75] evaluate on 14 Leeds Sports Pose (LSP)[67] keypoints, while others [43, 39, 17] evaluate on 17 Human3.6M keypoints, and some other works[76] do not explain clearly which keypoints they use. In Table 4.1, we report the results of both 14 and 17 keypoints for a fair comparison. We achieve SOTA performance, PA-MPJPE 40.3mm. with a single hypothesis.

Results on Human3.6M. Although ZeDO is a Zero-shot Diffusion-based Optimization pipeline, ZeDO achieves comparable results with learning-based. Following [26], we report the minMPJPE of multi-hypothesis, i.e., S number of initial poses, in inference and use the detected 2D poses as input. As shown in Table 4.2, on the Human3.6M dataset, we obtain 51.4mm in minMPJPE with $S = 50$, which is comparable with SOTA learning-based methods, while ZeDO does not train with any 2D-3D or image-3D pairs. Compared with other optimization-based methods, ZeDO outperforms previous works by a large margin, even with $S = 1$. In this experiment, we train the pose generation model on the training set of Human3.6M.

Results on 3DHP. Following previous works[43, 39, 17], we use ground truth 2D poses as input. As shown in Table 4.3, with $S = 50$, ZeDO even outperforms the learning-based methods by 2.7mm in minMPJPE. In the cross-domain evaluation, we achieve SOTA performance as 67.9mm in minMPJPE and outperform the optimization-based method by a large margin, with $S = 50$.

4.4 Discussion

4.4.1 Ablation Studies

Different diffusion-based pose generation models. As shown in table 4.5, we try to evaluate the cross-domain performance of our pipeline with different diffusion-based backbones on the 3DPW. We test the Score Matching Network[157], DDPM[56], and DDIM[151] models trained on the Human3.6M dataset and keep all other settings the same. It turns out that DDIM also achieves comparable performance in terms of PA-MPJPE and even lower MPJPE compared with the Score Matching Network we report above. The outcome validates the generality and viability of our idea, regardless of the specific structure of the diffusion backbone.

How does initial pose optimizer help ZeDO? The initial pose optimizer is designed to align the initial pose with the target 2D pose by rotation for better initialization. As shown in Table 4.4, the combination of rotation optimization and warmup iterations reduces the MPJPE by 9.3 mm when $S = 1$ and the minMPJPE by 2.0mm when $S = 50$, on the Human3.6M dataset. When $S = 50$, hypotheses cover lots of different pose orientations, resulting in the relatively smaller improved performance from the initial pose optimizer when S is larger. On the 3DHP dataset, the initial optimization further improves the performance by 34.5 mm in MPJPE when $S = 1$ since 3DHP contains more complex 3D human poses in different orientations than Human3.6M. The initial pose optimization is able to generate a reliable optimized initial pose and an optimal translation as the warmup translation for the following iterative optimization pipeline.

Does data augmentation help the performance? In ZeDO, the diffusion model is pre-trained for pose generation. However, the 3D pose distributions vary across different datasets. To ensure the pre-trained diffusion model can be adapted to different datasets, we utilize rotation and flip data augmentation during training. As expected, the data augmentation significantly improves the performance in cross-domain evaluation by 13.9 mm in MPJPE, shown in Table 4.4.

Boost the performance further by mixing dataset. According to Table 4.6, the pose generation model trained on the mixed datasets (Human3.6M + 3DHP) improves the performance of ZeDO by 1.3mm on Human3.6M and 2.8mm on 3DHP while using the same estimation algorithm.

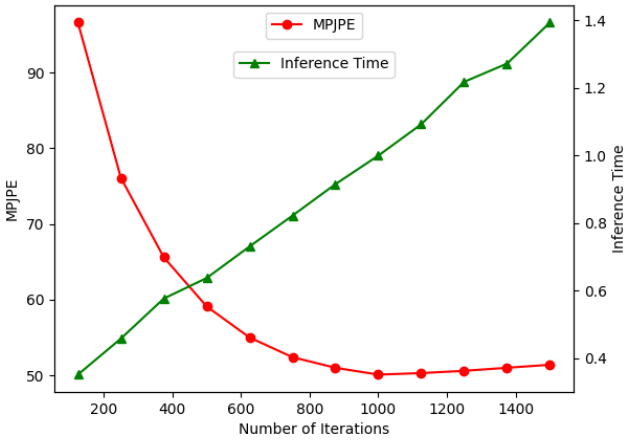


Figure 4.5: The MPJPE and inference time with different numbers of iterations on the Human3.6M dataset with ground truth 2D keypoints.

How to pick the initial pose? We utilize K-Means in our experiments to select anchor poses from the training set as initial poses. K-Means effectively finds the most representative poses in the training set, making it superior to other sampling strategies, as shown in Table 4.7. Random Sampling randomly samples a pose from the training set, whereas Random Generation is generated by the pre-trained pose generation model.

What is the best number of optimization iterations? In ZeDO, the number of diffusion optimization iterations is set to 1000. Intuitively, increasing the number of iterations can enhance performance but may suffer the inference speed. In Fig 4.5, as expected, the inference time increases linearly with respect to the number of iterations. However, the figure shows that the best performance is achieved when the number of iterations is around 1000. With the number of iterations exceeding 1000, there is no performance gain, and the inference speed decreases.

4.4.2 Limitations

Although ZeDO achieves state-of-the-art performance in various benchmarks and settings, several limitations still need to be further explored. 1) Similar to other optimization-based approaches, the optimizer in the loop requires camera intrinsic parameters. 2) Since our method is based on

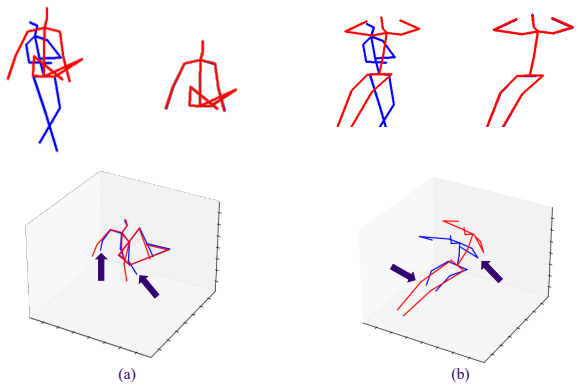


Figure 4.6: The failure cases of our method, because of the one-to-many issue in 3D HPE.

minimizing the 2D re-projection error, we are not able to solve the ambiguity of the depth and scale without additional information like bone length or height. 3) For an identical 2D human pose, there are multiple 3D human poses matched. Without image or temporal information, the 1-to-many mapping issue cannot be resolved by single frame lifting methods, as shown in Fig 4.6.

4.5 Conclusion

In this paper, we propose ZeDO, a Zero-shot Diffusion-based Optimization pipeline for 3D HPE. To the best of our knowledge, we are the first to introduce the diffusion model to the optimization-based method in the 3D HPE task. We leverage the pre-trained diffusion-based 3D human pose generation model and can optimize target 3D poses in the loop. To be specific, an optimizer that calculates the optimal translation is used iteratively with denoising steps in the diffusion model. Compared to other prior arts, ZeDO achieves state-of-the-art performance on Human3.6M, MPI-INF-3DHP, and 3DPW datasets, even with cross-dataset evaluation. In the future, we plan to further improve ZeDO by modifying the diffusion model and solving the limitations listed in Sec 4.4. It is our wish that this optimization method could become a common paradigm beyond end-to-end lifting networks in 3D human pose estimation tasks.

Algorithm 2 ZeDO pipeline

Require: Initial 3D pose P_{init}, Target 2D pose p_{2d},

2D pose confidence scores C_{2d}, Camera intrinsic K,

Diffusion timestamp t, Pre-trained diffusion model $\theta_g(P, t)$

$R_0, T_0 \leftarrow \arg\min_{R_0,T_0} \|K(R_0 P_{init} + T_0) - p_{2d}\|_2$

// Initial Pose Optimization

$P_0 \leftarrow R_0 P_{init}$

// Iterative Optimization and Denoising

$r \leftarrow K^{-1} p_{2d}$

$\hat{r} \leftarrow \frac{r}{\|r\|_2}$

for $i \leftarrow 0$ to $n-1$ **do**

  if $i < warmup$ **then**

    $T_i \leftarrow T_0$

  else

    $T_i \leftarrow \arg\min_{T_i} \|C_{2d}(K(P_i + T_i) - p_{2d})\|_2$

  end if

  // Project 3D keypoints to rays

  $\widetilde{P}_i \leftarrow \Big((P_i + T_i) \cdot \hat{r}\Big)\hat{r} - T_i$

  $P_{i+1} \leftarrow \theta_g(\widetilde{P}_i, t(i))$

end for

return P_n

Methods	CE	Opt	PA-MPJPE ↓	MPJPE ↓
Kolotouros *et al*[76]			59.2	96.9
Kocabas *et al*[74]			51.9	82.9
Kocabas *et al*[75]			46.4	74.7
Li *et al*[85]			<u>45.0</u>	<u>74.1</u>
Ma *et al*[109]			**41.3**	**67.5**
Li *et al*[85]	✓		50.9	82.0
Kocabas *et al*[74]	✓		56.5	93.5
Kocabas *et al*[75]	✓		50.9	82.0
Gong *et al*[43]	✓		58.5	94.1
Gholami *et al*[39]	✓		46.5	<u>81.2</u>
Chai *et al*[17]	✓		55.3	87.7
Song *et al*[153]	✓		55.9	-
Choutas *et al*[23]	✓		52.2	-
ZeDO $(S = 1, J = 17)$	✓	✓	**40.3**	**69.7**
ZeDO $(S = 1, J = 14)$	✓	✓	<u>43.1</u>	76.6

Table 4.1: Cross-domain evaluation results on 3DPW dataset. CE stands for cross-domain evaluation, and Opt means optimization-based method. Ground truth 2D poses are used.

Learning Methods	MPJPE ↓	PA-MPJPE ↓
Martinez *et al*[112]	62.9	47.7
Zhao *et al*[215]	57.6	-
Pavllo *et al*[129] ($f = 1$)	52.7	40.9
Li *et al*[91]	<u>47.1</u>	32.7
Gong *et al*[43]	50.2	39.1
Gong *et al*[42] ($f = 1$)	49.7	<u>31.6</u>
Ci *et al*[26] ($S = 1$)	51.0	-
Ci *et al*[26] ($S = 10$)	**45.1**	**30.5**
Optimization Methods	MPJPE ↓	PA-MPJPE ↓
Wang *et al*[185]	88.0	-
Bogo *et al*[7]	82.3	-
Li *et al*[87]	78.6	-
Gu *et al*[46]	77.2	-
Song *et al*[153]	-	56.4
ZeDO ($S = 1$)	65.7	49.0
ZeDO ($S = 10$)	<u>57.3</u>	<u>45.1</u>
ZeDO ($S = 50$)	**51.4**	**42.1**

Table 4.2: 3D HPE quantitative results on Human3.6M dataset. S indicates the number of hypotheses. All results are reported in millimeters (mm). The pose generation model is trained on Human3.6M. Detected 2D poses by Stacked Hourglass are used.

Methods	CE	Opt	MPJPE ↓	PCK ↑	AUC ↑
Mehta *et al*[114]			124.7	76.6	40.4
Martinez *et al*[112]			84.3	85.0	52.0
Pavllo *et al*[129] ($f = 1$)			86.6	-	-
Li *et al*[90] ($f = 9$)			58.0	93.8	63.3
Zhang *et al*[208] ($f = 1$)			<u>57.9</u>	94.2	63.8
ZeDO ($S = 1$)		✓	86.5	82.6	53.8
ZeDO ($S = 50$)		✓	**55.2**	93.0	65.6
Kanazawa *et al*[69]	✓		113.2	77.1	40.7
Ci *et al*[26]	✓		-	86.9	-
Gong *et al*[43]	✓		73.0	88.6	57.3
Gholami *et al*[39]	✓		68.3	90.2	59.0
Chai *et al*[17]	✓		**61.3**	92.1	62.5
Müller *et al*[118]	✓	✓	101.2	-	-
ZeDO ($S = 1$)	✓	✓	99.9	81.8	50.9
ZeDO ($S = 50$)	✓	✓	<u>69.9</u>	90.2	58.8

Table 4.3: 3D HPE quantitative results on 3DHP dataset. CE stands for cross-domain evaluation, and Opt means optimization-based method. Ground truth 2D poses are used.

Dataset	Diff Model	RO	WU	RA	GT	$S = 1$		$S = 50$	
						MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
H36M	H36M					75.0	52.7	53.4	42.7
H36M	H36M	✓				77.2	53.7	52.7	42.4
H36M	H36M	✓	✓			**65.7** (9.3 ↓)	**49.0** (3.7 ↓)	**51.4** (2.0 ↓)	**42.1** (0.6 ↓)
H36M	H36M	✓	✓	✓		69.5	51.4	52.9	42.5
H36M	H36M	✓	✓		✓	50.1	35.8	37.0	27.5
3DHP	H36M				✓	148.3	88.8	93.4	59.0
3DHP	H36M	✓	✓		✓	113.8	74.1	80.1	56.0
3DHP	H36M	✓	✓	✓	✓	**99.9** (48.4 ↓)	**67.9** (20.9 ↓)	**69.9** (23.5 ↓)	**49.0** (10.0 ↓)
3DHP	3DHP	✓	✓		✓	86.5	55.9	55.2	38.6

Table 4.4: The ablation study results of ZeDO. RO stands for rotation optimization as the initial pose optimization. WU denotes the warmup iterations. RA is the rotation data augmentation for training the pose generation model. The dataset name under the Dataset column is the testing dataset, and the name under the Diff Model column is the dataset used for diffusion model pre-training.

Diffusion Backbone	PA-MPJPE ↓	MPJPE ↓
Score Matching [157]	**40.3**	<u>69.7</u>
DDIM[151]	<u>40.4</u>	**67.9**
DDPM[121]	51.7	81.3

Table 4.5: Different diffusion backbone 3D HPE quantitative results on 3DPW dataset.

Dataset	Diff Model	MPJPE ↓	PA-MPJPE ↓
H36M	H36M	37.0	27.5
H36M	mixed	**35.7**	**26.5**
3DHP	H36M	69.9	49.0
3DHP	3DHP	55.2	38.6
3DHP	mixed	**52.4**	**37.7**

Table 4.6: The pose generation models trained on mixed datasets (Human3.6M + 3DHP) achieves the best performance with $S = 50$, as well as better generalization. GT 2D keypoints are used.

Sampling	MPJPE ↓	PA-MPJPE ↓
Random Sampling	78.2	51.2
Random Generation	70.4	46.0
K-Means	**50.1**	**35.8**

Table 4.7: Results on Human3.6M dataset with different sampling strategies when $S = 1$. Ground truth 2D keypoints are used.

Chapter 5

UNIHPR: UNIFIED HUMAN POSE REPRESENTATION VIA SINGULAR VALUE CONTRASTIVE LEARNING

5.1 Introduction

As an important component of human-centric applications, human pose representations (HPRs) are critical in many downstream tasks, such as human pose estimation, action recognition, human-computer interaction, object tracking, etc. Recently, aligning text and human pose sequences (human motion)[171, 211, 172] has been widely discovered. However, there are many more data representations that can be used to denote human poses, including images, 2D keypoints, 3D skeletons, mesh models and *etc*. From the perspective of representation learning, many previous methods have been dedicated to mapping the representation of human pose sequences into the corresponding text space [171, 211, 172]. On the other hand, in this paper, we propose **UniHPR**, a Unified Human Pose Representation learning framework, which aims to align RGB images, 2D and 3D human poses in the shared feature space. In order to evaluate the quality of the proposed learned representation, we choose human pose estimation (HPE) as our evaluation task. By conducting task-specific fine-tuning, UniHPR can achieve the SOTA performance on both 2D and 3D HPE tasks.

Estimating 2D and 3D human poses (*i.e*, human keypoints) [7, 2, 159, 129, 91, 44, 17, 65] from only RGB images is one of the foundational tasks in the computer vision field, which can be further used for several downstream tasks like multiple object tracking [2, 147], action recognition [144, 32], human-computer interaction [189, 60], human body reconstruction [192, 84], sports application [217, 64], *etc*. Previous works follow the paradigms which estimate 3D human poses from 2D human poses (so-called lifting) [112, 129, 215, 220, 44] or directly regress 3D human poses from RGB images (image-based) [77, 37, 161, 163, 206, 205]. Lifting networks learn the

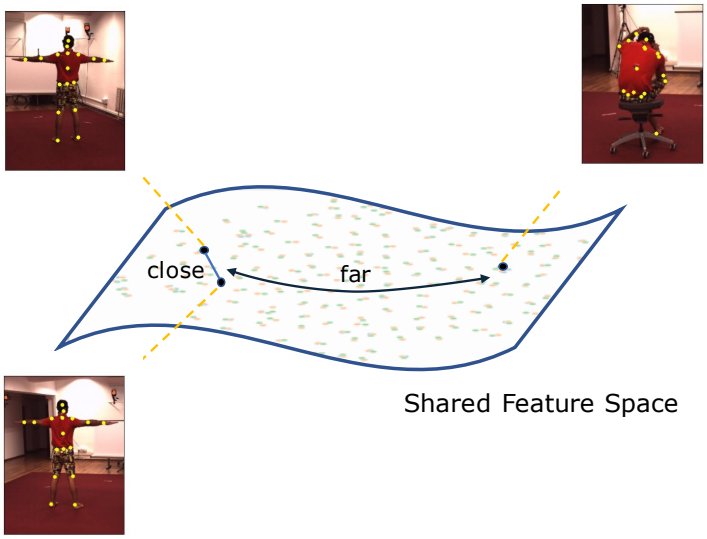


Figure 5.1: RGB image, 2D and 3D human pose embeddings extracted by corresponding encoders in the shared feature space. After conducting contrastive learning during the pre-training stage, the embeddings extracted from these three different data representations of the same training sample are close to each other and away from other negative samples.

mapping between 2D and 3D human poses, and the image-based methods take advantage of the rich image information to directly get accurate 3D pose estimation results.

Learning joint embeddings across more than two data representations (or modalities) is quite challenging. Inspired by Contrastive Language-Image Pre-Training (CLIP) [134], which proposes to learn aligned visual features with natural language supervisions trained on web-scale image-text paired data. We claim that alignment among RGB image, 2D and 3D human pose representations can also benefit from contrastive learning on large-scale and diverse datasets (*e.g*, Human3.6M [61], MPI-INF-3DHP [113], *etc*).

During the evaluation, UniHPR serves as an encoder with additional downstream task decoders for 2D or 3D HPE. Therefore, the whole pipeline consists of image, 2D and 3D human pose encoders, and 2D and 3D human pose decoders. The embedding features of these three data representations are aligned and shared. To be specific, we first encode the images by HRNet [188],

2D and 3D human poses by shallow Transformers [180] respectively to get the corresponding embeddings, respectively. We then conduct contrastive learning to align the embeddings from these three different data representations of the same training sample in the shared feature space for the unified representation learning. However, aligning embeddings from more than two data representations is challenging, and therefore, we propose a singular value based supervised contrastive learning loss to align three data representations at the same time. After that, during the training stage, we jointly train encoders and decoders with contrastive learning and multi-task learning simultaneously. During inference, since the embeddings are aligned in the same feature space, UniHPR can simultaneously support 2D human pose estimation and 3D human pose estimation, both lifting-based and image-based, in the same pipeline.

Our contributions can be summarised as follows:

- We propose the singular value based InfoNCE loss for supervised contrastive learning to effectively align embedding of more than two data representations at the same time.

- UniHPR aligns the embedding of Human Pose Representations from three distinctive data representations, i.e., images, 2D and 3D human poses.

- With a simple additional diffusion-based decoder, UniHPR achieves SOTA performance on frame-based 3D HPE tasks, e.g., MPJPE 49.9mm on the Human3.6M dataset with image-3D branch and PA-MPJPE 51.6mm with 2D-3D branch on the 3DPW dataset for the 3D human pose estimation task.

5.2 Methodology

We build a unified human pose representation learning pipeline. During training, for any triplet of the cropped human image, $I \in \mathbb{N}^{H \times W \times 3}$, 2D and 3D human poses, $P_{2D/3D} \in \mathbb{R}^{J \times 2/3}$, UniHPR aligns the embeddings from all three representations and utilizes 2D and 3D pose decoders for downstream tasks.

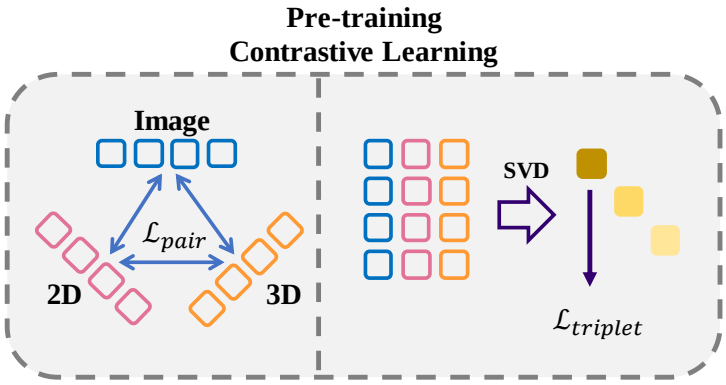


Figure 5.2: $\mathcal{L}_{pair}$ is applied three times for contrastive learning and the singular value based $\mathcal{L}_{triplet}$ focuses on aligning three representations at the same time.

5.2.1 Framework Architecture

Image encoder. The extraction of embedding from an RGB image is based on the HRNet [188], which is a convolution-based backbone for various visual recognition tasks. We concatenate and flatten the average pooled features from the last stage and pass it through a linear projection layer to obtain a 1-D embedding as our image representation.

2D/3D pose encoders. We adopt two Transformer-based [180] encoders to extract the embeddings from 2D and 3D human poses, respectively. We conduct bounding box normalized keypoint-wise patch embedding and retain the spatial information of each keypoint via adding learnable spatial position embedding. Then, the pose tokens prepended with a $[CLS]$ token and a bounding box token are fed into standard transformer encoder layers, including multi-head self-attention, feed-forward layers, and normalization layers. After that, we use the $[CLS]$ tokens as 2D and 3D pose embedding, respectively, which effectively aggregates the information of the other tokens and can be regarded as general prior.

2D and 3D pose decoders. We try several different architectures for our task specific decoder, including an MLP, a transformer and a diffusion based model. The diffusion decoder provides the

Algorithm 3 Random Sampling in $\mathcal{L}_{svd}$

Require: Image embedding, x_{img}, 2D pose embedding, x_{2D}, 3D pose embedding, x_{3D}, Batch size, B

index_list $\leftarrow$ zeros$(B, B, 3)$
for $i \leftarrow 0, B$ **do**
  index_list $[:, i, 0] \leftarrow$ arange(B)
  index_list $[:, 0, i] \leftarrow$ arange(B)
end for
for $i \leftarrow 1, B$ **do**
  for $j \leftarrow 1, 3$ **do**
    index_list $[:, i, j] \leftarrow$ shuffle(arange(B))
  end for
end for

Algorithm 4 Implementation of $\mathcal{L}_{svd}$

Require: Normalized image embedding, x_{img} 2D pose embedding, x_{2D}, 3D pose embedding, x_{3D}, Batch size, B, Temperature, τ

$\mathcal{M}_x \leftarrow$ stack$((x_{img}, x_{2d}, x_{3d}), dim = 1)$
index_list $\leftarrow$ RandSample$(x_{img}, x_{2D}, x_{3D})$
\# $\mathcal{M}_x \in \mathbb{R}^{B \times B \times 3 \times D}$
$\mathcal{M}_x \leftarrow \mathcal{M}_x$ [index_list]
\# $\mathcal{M}_x \in \mathbb{R}^{B \times B \times 3 \times 3}$
$\mathcal{M}_x \leftarrow \mathcal{M}_x \mathcal{M}_x^{\intercal}$
logits $\leftarrow$ eigenval$(\mathcal{M}_x)$ $[:, :, 0]$
logits $\leftarrow$ logits$/\tau$
label $\leftarrow$ zeros(B)
$\mathcal{L}_{triplet} =$ CrossEntropy(logits, label)

Algorithm 5.1: The **pseudo-code** of triplet random sampling and the implementation of $\mathcal{L}_{triplet}$. To simplify the computation, for each mini batch, we randomly sample $B - 1$ negative triplets and one positive triplet.

best results in decoding the embedding to generate 2D and 3D human poses. We treat the decoders following the Score Matching paradigm [156], instead of using DDPM[57] or DDIM[152]. To be specific, the encoded embedding is added with time embedding as well as a data representation token, which indicates the source of the embedding (*e.g*from an image, 2D or 3D pose) in the diffusion network as a condition embedding and is used to generate the final 2D and 3D poses. The detailed architectures of all decoders are in the supplemental material.

5.2.2 Unified Representation Learning via Contrastive Learning

During the representation learning stage, we aim to align the embeddings from images, 2D and 3D human poses via the supervised contrastive learning. Given a batch of data, we have the RGB images, $I \in \mathbb{N}^{B \times H \times W \times 3}$, 2D poses, $P_{2D} \in \mathbb{R}^{B \times J \times 2}$, and 3D poses, $P_{3D} \in \mathbb{R}^{B \times J \times 3}$, where B, H, W, J are batch size, image height and width, and number of human body keypoints, re-

spectively. The image, 2D, and 3D pose encoders E_{img}, E_{2D}, E_{3D} are trained by maximizing the similarity between image embedding $x_{img} \in \mathbb{R}^{B\times D}$, 2D pose embedding $x_{2D} \in \mathbb{R}^{B\times D}$, and 3D pose embedding $x_{3D} \in \mathbb{R}^{B\times D}$, where D is the dimension of the embedding, which is the same over all three data representations. The most intuitive approach to aligning three embeddings is to apply three pair-wise contrastive losses. For embeddings, $x_{\mathcal{S}}, x_{\mathcal{T}}$, from any pair of data representations, the contrastive learning loss is

$$\mathcal{L}_{pair} = -\log \frac{\exp\left(x_{\mathcal{S}} \cdot x_{\mathcal{T}}^{+}/\tau\right)}{\sum_{i=1}^{B} \exp(x_{\mathcal{S}} \cdot x_{\mathcal{T},i}/\tau)}, \tag{5.1}$$

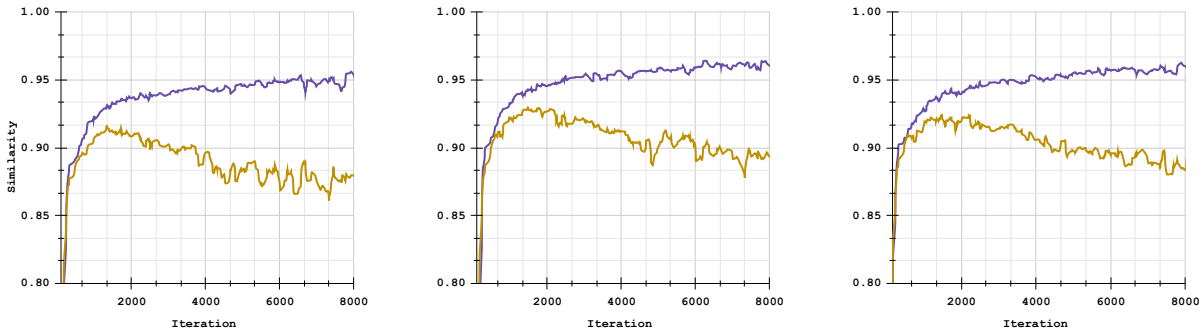
where τ is the learnable temperature initialized by τ_0.

However, we found that simply applying three pairwise InfoNCE loss cannot obtain expected embedding similarity across three representations, as shown in the ablation studies in Section 5.3.5. Therefore, we propose a singular value-based InfoNCE loss (Triplet-InfoNCE) to address this issue.

We stack the embeddings from three representations to build a normalized embedding matrix, formulated by

$$\mathcal{M}_x = \begin{bmatrix} x_{img} & x_{2D} & x_{3D} \end{bmatrix}^T \in \mathbb{R}^{3\times D}. \tag{5.2}$$

If we apply singular value decomposition (SVD) to this matrix, $M_x = U\Sigma V^*$, the largest singular value, $\sigma_1 = \Sigma_{11}$, is related to the linear correlation of row vectors. Meanwhile, since the embeddings are normalized, the largest singular value should be in $\left[-\sqrt{3}, \sqrt{3}\right]$. Therefore, we can use InfoNCE loss to align any triplet of embeddings by maximizing the σ_1. However, computing the singular value of a matrix with $3 \times D$, where $3 \ll D$, is time-consuming. Therefore, to accelerate the training procedure, instead of σ_1, the largest eigenvalue λ_1 of the matrix $\mathcal{M}_x\mathcal{M}_x^{\intercal} \in \mathbb{R}^{3\times 3}$ is the optimization target, since $\lambda_1 = \sigma_1^2$. Therefore, by maximizing the λ_1 for positive triplets, which contain three embeddings from the same frame, and minimizing the λ_1 for negative triplets, which contain at least one embedding from a different frame, we are able to align embeddings from three representations jointly.



(a) 2D-3D poe (b) image-2D pose (c) image-3D pose

Figure 5.2: **Cosine similarities** between different data representations. The yellow line is the one trained only with three pair-wise losses, $\mathcal{L}_{pair}$, and the purple line is the training curve with additional singular value-based InfoNCE loss, $\mathcal{L}_{triplet}$. Our proposed singular value-based InfoNCE loss helps align the feature space.

However, in one minibatch, the number of negative triplets for any positive triplet is $3B^2 - 3B + 1$, and if we use all the negative samples as our denominator in InfoNCE loss, the time consumption is unacceptable. As shown in the Alg 5.1, we apply a random sample algorithm to select only $B - 1$ negative triplets for each positive triplet. In this case, the singular value based InfoNCE loss can be formulated as,

$$\mathcal{L}_{triplet} = -\log \frac{\exp\left(\lambda_1^+/\tau\right)}{\sum_{i=1}^{B} \exp(\lambda_{1i}/\tau)}. \tag{5.3}$$

Overall, our contrastive learning loss is

$$\mathcal{L}_{cl} = \mathcal{L}_{pair} + \alpha \mathcal{L}_{triplet}. \tag{5.4}$$

where α is the weighted factor.

	Method	3DPW	Human3.6M	
		PA-MPJPE (↓)	MPJPE (↓)	PA-MPJPE (↓)
Temporal	VideoPose3D (f=243) [129]	68.0	46.8	36.5
	AdaptPose† [39]	46.5	-	-
	Li *et al* [88]	-	43.7	35.2
	MixSTE [208]	-	40.9	32.6
	MPM [214]	-	42.6	34.7
Frame-based	SimpleBaseline† [112]	89.4	62.9	47.7
	SemGCN† [215]	102.0	61.2	47.7
	VideoPose3D† (f=1) [129]	94.6	55.2	42.3
	PoseAug† [44]	58.5	<u>52.9</u>	-
	PoseDA† [17]	<u>55.3</u>	-	-
	UniHPR† (ours)	**51.6**	**52.6**	**39.9**

Table 5.1: **Lifting-based 3D HPE** performance on the 3DPW and Human3.6M datasets under MPJPE and PA-MPJPE. The ground truth 2D keypoints are used on 3DPW dataset, while the detected 2D keypoints from CPN are used on Human3.6M dataset. † indicates cross-domain evaluation on 3DPW dataset.

5.2.3 Task-Specific Finetune

After the representation learning stage, all encoders and decoders are trained jointly. While encoders are trained with $\mathcal{L}_{cl}$, the task losses, $\mathcal{L}_{2D/3D}$, depend on the architectures of decoders. For the diffusion-based decoder, we adopt the loss from the Score Matching Network [157], and for the MLP-based decoder, we utilize $L2$ loss.

Therefore, the overall loss in Task-Specific Finetune is

Retrieval	MPJPE (↓)	PA-MPJPE (↓)
2D-3D	9.2	7.1
Image-3D	10.4	7.6

Table 5.2: **Quantitative evaluation** of the unified representation. Pose retrieval on Human3.6M test dataset.

Retrieval	Top-1 Acc. (↑)	Top-3 Acc. (↑)
3D-Image	89.2	95.6
2D-Image	95.5	97.6

Table 5.3: **Quantitative evaluation** of the unified representation. Image retrieval on Human3.6M test dataset with 1 FPS.

$$\mathcal{L} = \mathcal{L}_{cl} + \mathcal{L}_{2D} + \mathcal{L}_{3D}. \tag{5.5}$$

During inference, since the embeddings are well-aligned unified human pose representations in the same feature space, UniHPR can utilize the embedding from any representation and estimate 2D or 3D human poses with shared decoders.

5.3 Experiments

5.3.1 Implementation Details

We implement our proposed framework using PyTorch [126] on a single NVIDIA A100/80G GPU. The representation learning includes two steps: (1) 2D-3D alignment; (2) Image-2D-3D joint alignment (see Fig. 2); followed by a task-specific finetuning stage. In the first step of representation learning, the batch size is 2048, $\tau_0 = 1/14$, and $\tau \in [1/100, 10^4]$, while in the second step, the batch size is 180, $\tau_0 = 1/5$, and $\tau \in [1/10, 10^4]$. During the multi-task training steps, encoders and decoders are trained together with the batch size being 180, $\tau_0 = 1/5$, and $\tau \in [1/5, 10^4]$. For the weight of triplet contrastive loss, $\mathcal{L}_{triplet}$, $\alpha = 1$. The input image size of the image encoder is 192×256. During both of the two steps, we adopt Adam optimizer with a learning rate of 1×10^{-4}. We train UniHPR on Human3.6M [61] and MPI-INF-3DHP [113] datasets and apply ablation study about the performance difference on different training datasets.

5.3.2 Datasets and Performance Metrics

To conduct the quantitative performance evaluation of the proposed UniHPR, we use several widely used 3D human pose datasets to train and evaluate our proposed framework, including Human3.6M [61], MPI-INF-3DHP [113], and 3DPW [182]. We train UniHPR on Human3.6M and MPI-INI-3DHP and evaluate it on Human3.6M and 3DPW.

Human3.6M[61] dataset, which contains 3.6 million frames of corresponding 2D and 3D human poses, including 5 female and 6 male subjects under 17 different scenarios, is a video dataset captured using a MoCap system. Following the previous works for fair comparisons, we choose 5 subjects (S1, S5, S6, S7, S8) for training, and the other 2 subjects (S9 and S11) for evaluation. We report the Mean Per Joint Position Error (MPJPE) as the performance metric of Protocol #1 as well as Procrusts analysis MPJPE (PA-MPJPE) as the metric of Protocol #2 for both the lifting path and image-bath path in our proposed framework.

MPI-INF-3DHP[113] is a more challenging 3D human pose dataset, which is captured both indoors and outdoors, while Human3.6M is captured only indoors. MPI-INF-3DHP contains 1.3M frames, consisting of 4 males and 4 females with 8 types of action captured by 14 cameras covering a greater diversity of poses. We use MPI-INF-3DHP in training to enrich the training samples and enhance the performance of our image branch.

3DPW[182] is the first dataset that includes video footage taken from a moving phone camera. It includes 60 video sequences as the training dataset with 22k images, while the testing dataset includes 35k images. Compared to Human 3.6M or MPI-INF-3DHP, 3DPW is an even more challenging in-the-wild dataset, with uncontrolled motion and scene. Similar to most other works, we only evaluate our model under PA-MPJPE metrics, without considering MPJPE, due to the fact that the scale of the human body, camera intrinsic, and distance of 3DPW are not compatible with the training data.

5.3.3 Evaluation of the Unified Human Pose Representation

Quantitative Evaluation of Representation Learning. To better evaluate the quality of learned unified representations, we conduct Pose and Image Retrieval on Human3.6M dataset. The retrieved 3D human pose or image has the most similar 3D pose or image embedding with the image, 2D or 3D pose representation query. For Image Retrieval task, the FPS is set as 1. In Table 5.3, 2D-3D Pose Retrieval can achieve MPJPE 9.2mm and the MPJPE of Image-3D Pose Retrieval is 10.4mm, and 2D-Image Image Retrieval can achieve Top-1 Accuracy 95.5%, which illustrate the unified representations are well aligned in images, 2D and 3D human poses. More visualization is included in the supplementary material.

5.3.4 Evaluation of Human Pose Estimation

Lifting-based 3D Human Pose Estimation We evaluate the performance of lifting-based 3D HPE tasks on Human3.6M and 3DPW datasets. As shown in Table 5.1, UniHPR archives 51.6 mm in terms of PA-MPJPE on 3DPW dataset and 52.6 mm in terms of MPJPE on Human3.6M dataset, which is the state-of-the-art performance. Since UniHPR is not trained on 3DPW, it is a fair comparison with those cross-domain evaluation methods.

Image-based 3D Human Pose Estimation As for image-based 3D HPE, we also evaluate the performance on Human3.6M and 3DPW datasets. As shown in Table 5.4, UniHPR respectively achieves 49.9 mm and 35.7 mm in terms of MPJPE and PA-MPJPE on Human3.6M dataset, as well as 65.7 mm of PA-MPJPE on 3DPW dataset. Note that we are the only keypoint-based method in Table 5.4, and all the others are SMPL-based. UniHPR achieves comparable performance regarding the number of model parameters and training data with SOTA methods.

5.3.5 Ablation Study

In this section, we conduct extensive ablation studies to investigate the importance of each module in the UniHPR, especially how our proposed singular value based loss, $\mathcal{L}_{triple}$, helps the training and improves the performance.

End-to-End training without alignment. We claim that feature alignment, i.e., pre-training via contrastive learning, among different representations is the key to success. Therefore, we conduct the ablation studies on skipping the alignment training stages. As shown in Table 5.5, alignment improves the image-based 3D HPE performance significantly on the Human3.6M dataset. As shown in table 5.5, without the 2-step contrastive learning, the performance gap between lifting and image branches shows that the features are not correctly aligned. Furthermore, the combination of $\mathcal{L}_{triplet}$ and $\mathcal{L}_{pair}$ provides the best performance on both lifting and image branches, which is consistent with the comparative results shown in Fig 5.2.

Ablation on representation token, $\mathcal{R}$. In UniHPR, we design a representation token when using the 3D pose decoder to estimate 3D human poses. The representation token indicates which representation the features derived from either (*e.g.*image or 2D pose). As shown in Table 5.5, consistent improvement is observed in using the representation token among lifting-based and image-based 3D HPE tasks on the Human3.6M dataset.

Effectiveness of the $\mathcal{L}_{triplet}$. As shown in Figure 5.2, compared to simply applying three pairwise InfoNCE loss, $\mathcal{L}_{pair}$, the proposed singular value-based InfoNCE loss, $\mathcal{L}_{triplet}$, significantly better aligns the features from different representations. With the help of $\mathcal{L}_{triplet}$, the embedding cosine similarity between different representation does not decrease after around 1500 iterations and keeps increasing to around 0.95 in 8000 iterations. For quantitative evaluation, in Table 5.5, without the help of $\mathcal{L}_{triplet}$, three $\mathcal{L}_{pair}$ can only achieve MPJPE 65.5mm and 60.0mm for image and keypoint branches, which are 6.8mm and 19.1mm more than the jointly trained model.

Training with additional data. As shown in Table 5.5, it is noted that the distribution of 2D and 3D pose pairs on 3DHP differs from Human3.6M, which increases the robustness of the lifting-based branch but decreases the performance slightly on Human3.6M, since the model trained with both Human3.6M and 3DHP achieves the best performance on 3DPW. Furthermore, training with additional data boosts the image-based branch by improving the diversity of image data.

5.4 Conclusion

In conclusion, the UniHPR framework represents a significant step forward in unified human pose representation learning by mitigating the gap between image, 2D and 3D human pose representations. Despite its potential limitations in data and computational requirements, UniHPR sets a promising direction for future research, particularly in improving generalization capabilities and multi-modal representation learning. The framework's achievements on benchmark datasets like Human3.6M and 3DPW justify its potential, paving the way for advancements in applications across multiple domains such as text-to-pose and pose-to-image generation.

	Method	Representation	3DPW	Human3.6M	
			PA-MPJPE (↓)	MPJPE (↓)	PA-MPJPE (↓)
Temporal	Kanazawa *et al* [70]	SMPL	72.6	-	56.9
	Doersch *et al* [30]	SMPL	74.7	-	-
	Arnab *et al* [3]	SMPL	72.2	77.8	54.3
	DSD [164]	SMPL	69.5	59.1	42.4
	VIBE [74]	SMPL	56.5	65.9	41.5
Frame-based	Pavlakos *et al* [128]	SMPL	-	-	75.9
	HMR [69]	SMPL	76.7	88.0	56.8
	NBF [122]	SMPL	-	-	59.9
	GraphCMR [79]	SMPL	70.2	-	50.1
	HoloPose [48]	SMPL	-	60.3	46.5
	DenseRaC [196]	SMPL	-	76.8	48.0
	SPIN [77]	SMPL	59.2	62.5	41.1
	DecoMR [203]	SMPL	61.7	-	39.3
	HKMR [37]	SMPL	-	59.6	43.2
	PyMAF [206]	SMPL	58.9	57.7	40.5
	PARE [75]	SMPL	50.9	76.8	50.6
	PyMAF-X [205]	SMPL	<u>47.1</u>	54.2	37.2
	CLIFF [91]	SMPL	**43.0**	**47.1**	**32.7**
	UniHPR-w32† (ours)	Keypoint	65.8	54.5	39.5
	UniHPR-w48† (ours)	Keypoint	64.5	<u>49.9</u>	<u>35.7</u>

Table 5.4: **Image-based 3D HPE** performance on the 3DPW and Human3.6M datasets under MPJPE and PA-MPJPE. † indicates cross-domain evaluation on 3DPW dataset.

				GT 2D		Image	
$\mathcal{L}_{pair}$	$\mathcal{L}_{triplet}$	$\mathcal{R}$	w. 3DHP	MPJPE (↓)	PA-MPJPE (↓)	MPJPE (↓)	PA-MPJPE (↓)
baseline				41.3	31.6	91.8	68.7
✓				60.0 (+18.7)	47.5 (+15.9)	65.5 (-26.3)	51.8 (-16.9)
✓	✓			40.9 (-0.4)	31.7 (+0.1)	58.7 (-33.1)	44.4 (-24.3)
✓	✓	✓		**39.3 (-2.0)**	**29.9 (-1.7)**	57.5 (-34.3)	42.9 (-25.8)
✓	✓	✓	✓	41.7 (+0.4)	32.6 (+1.0)	**54.5 (-37.3)**	**39.5 (-29.2)**

Table 5.5: **Ablation study** on UniHPR. Evaluated on Human3.6M dataset. $\mathcal{L}_{pair}$ and $\mathcal{L}_{triplet}$ denotes applying those losses on the pre-training stage. $\mathcal{M}$ token means decoders utilize the representation token. We evaluate the performance with additional data from MPI-INF-3DHP dataset as well.

Chapter 6

PACKDIT: JOINT HUMAN MOTION AND TEXT GENERATION VIA MUTUAL PROMPTING

6.1 Introduction

Human motion capture is widely used across multiple industries, including film production, video game development, and virtual reality (VR). However, setting up a motion capture studio is expensive, and the quality of the captured motion heavily depends on the actors' performance. With advancements in diffusion models [55, 150, 154, 157], recent years have seen substantial progress in Motion Generation [226, 174, 51, 211, 63, 212, 207, 176], which aims to automatically generate rich, realistic human motion sequences.

Motion generation encompasses the production of human motion sequences with or without conditions from other modalities, such as action classes, text, audio, music, and speech. Among these, text stands out for its ability to convey detailed information about actions, speeds, directions, and goals, either explicitly or implicitly. For instance, HumanML3D [50] is a comprehensive text-to-motion generation dataset that provides well-annotated text-motion pairs derived from HumanAct12 [52] and AMASS [110]. Recent studies leverage such datasets to explore diffusion-based models for text-to-motion generation and autoregressive models for motion-to-text understanding.

However, few methods can do both text-to-motion and motion-to-text generation. Recently, MotionGPT [63] uses the auto-regressive paradigm to achieve this goal. Addressing the challenges of effectively generating and integrating motion and text, we propose a novel framework, **PackDiT**, the first diffusion-based text-motion joint generation model. PackDiT stands out for its flexibility and capability to handle multiple tasks within a unified architecture, leveraging two independent Diffusion Transformers (DiTs), Motion DiT and Text DiT, with mutual blocks and multi-stage training strategies. PackDiT is initially pre-trained unconditionally, then jointly trained and fine-

Table 6.1: Comparison of recent state-of-the-art methods on diverse motion-relevant tasks. *Random Motion* and *Random Text* represent unconditional generation of motions and motion descriptions. *Joint Gen* means the joint generation of motion and motion descriptions.

Methods	Text-to-Motion	Motion-to-Text	Motion Pred.	Motion In-Between	Random Motion	Random Text	Joint Gen.
T2M-GPT [207]	✓	-	-	-	✓	-	-
MLD [20]	✓	-	-	-	✓	-	-
TM2T [51]	✓	✓	-	-	-	-	-
MDM [174]	✓	-	✓	✓	✓	-	-
MotionDiffuse [211]	✓	-	✓	✓	✓	-	-
LMM [212]	✓	-	✓	✓	✓	-	-
MotionGPT [63]	✓	✓	✓	✓	✓	✓	-
PackDiT (ours)	✓	✓	✓	✓	✓	✓	✓

tuned, enhancing fidelity and alignment.

We evaluate PackDiT on the HumanML3D dataset [50] across a range of tasks and corresponding metrics. Compared to other state-of-the-art text-to-motion generative models, PackDiT achieves superior performance on the FID metric with fewer parameters. Additionally, PackDiT demonstrates leading performance in motion prediction and in-between tasks. Notably, we are the first to show that a diffusion-based generative model can perform motion-to-text generation, achieving comparable results to large language models (LLMs) trained on extensive text corpora.

In particular, we make the following contributions:

- We are the first diffusion-based model that can accomplish diverse motion-relevant tasks, including text-to-motion, motion-to-text, and motion-text joint generation, *etc*.
- We show that by adding mutual blocks between text and motion diffusion generative models (*e.g*DiT), we can easily package two separated models to achieve good joint generation ability.
- Our experiments show that our proposed method achieves state-of-the-art text-to-motion performance with FID as 0.106, as well as the motion prediction and motion in-between tasks.

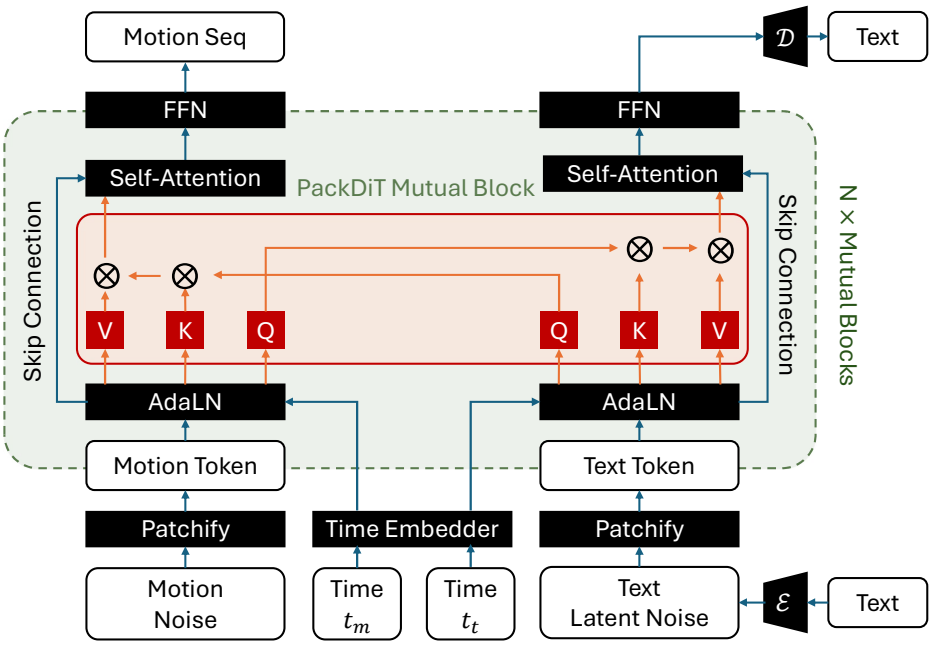


Figure 6.1: The architecture of PackDiT, where there are two independent DiTs for Motion and Text generation. By enabling and disabling the cross-attention layers in-between, PackDiT can solve almost all motion and text-related generation tasks, including text-to-motion, motion-to-text, motion prediction, motion in-between, random motion and text generation, and joint motion-text generation.

6.2 Method

To simulatneously tackle the motion-to-text, text-to-motion, and joint generation issues, we propose PackDiT, which is a flexible motion and text generation pipeline and consists of two DiT models, i.e., Motion DiT, $D_{\mathcal{M}}$ and Text DiT, $D_{\mathcal{T}}$. As shown in Table 6.1, compared with previous works based on diffusion models or large language models, PackDiT is able to achieve the most motion- and text-related generation tasks. For motion representations, we follow [50, 63] and represent motion of frame i as $\mathcal{M}_i \in \mathbb{R}^{263}$, where $\mathcal{M}_i = \{R_i, h_i^r, v_i^r, J_r, J_p, J_v, F\}$. R_i is the global rotation of the human body, h_i^r is the height of the root point, v_i^r is the velocity of the root

point, J_r is the relative rotation of each of joints, J_p is the position of each joint in canonical view, J_v is the velocity of each joint, and F represent each foot is on the ground or now. Following UniDiffuser [4], a text encoder, BERT [29], and a text decoder, GPT-2 [135], are utilized for text generation.

6.2.1 Unconditional Motion Generation

To build a flexible multi-task joint generation diffusion pipeline, all inputs and outputs should be formulated into tokens, which allow us able to integrate mutual blocks for information exchange between Text and Motion. Therefore, we adopt DiT [130] as our baseline diffusion model. As shown on the left of Fig 6.1, for each timestep, t_i, motion representations $\mathcal{M} = \{\mathcal{M}_0, \ldots, \mathcal{M}_n\}$ are converted to motion tokens after Patchify. Then, the Motion DiT, $D_{\mathcal{M}}$, generates the noise. The training procedure follows DDIM [150]. Therefore, the training loss of $D_{\mathcal{M}}$ is

$$\mathcal{L}_{\mathcal{M}} = \|\epsilon - D_{\mathcal{M}}(\sqrt{\overline{\alpha}_t}\mathcal{M} + \sqrt{1-\overline{\alpha}_t}\epsilon, t)\|_2^2. \tag{6.1}$$

6.2.2 Unconditional Text Generation

Following UniDiffuser [4], to make the diffusion model generate texts, we utilize a text encoder, T_{enc}, and a text decoder, T_{dec}, for better text generation quality. As shown on the right of Fig 6.1, similar to our motion diffusion model, after encoder and forward diffusion, the text latent noise is fed to Text DiT, $D_{\mathcal{T}}$. After the reverse diffusion, all generated text tokens are treated as the prefix of the text decoder and used to generate the corresponding texts. Following DDIM [150], the training objective of $D_{\mathcal{T}}$ is

$$\mathcal{L}_{\mathcal{T}} = \|\epsilon - D_{\mathcal{T}}(\sqrt{\overline{\alpha}_t}\mathcal{T} + \sqrt{1-\overline{\alpha}_t}\epsilon, t)\|_2^2. \tag{6.2}$$

However, the diffusion model struggles to directly generate text tokens because they are discrete and have relatively high dimensions. To reduce the dimensionality of the encoded text tokens from the text encoder and create more continuous text representations, we further train an additional

projection model, P, which projects the encoded text tokens to the text latent tokens, $\mathcal{T}$, used in $D_{\mathcal{T}}$, and re-projects the generated text tokens to the text decoder.

6.2.3 Mutual Prompting

Compared with other joint generation pipelines [165, 4, 142, 169], our pipeline does not require training a unified generation model for all modalities. Instead, the generation models for each modality operate independently and are integrated using mutual blocks. This architectural choice makes PackDiT significantly more flexible and capable of handling joint generation tasks more efficiently. Each modality-specific model can be optimized independently, allowing for specialized fine-tuning and improvements without affecting the entire system. This flexibility extends to scaling the system for new modalities, as new generation models can be added without extensive reconfiguration of the existing pipeline. As a result, PackDiT offers a robust and adaptable diverse joint generation.

During training or inference, the intermediate tokens, M and T, from $D_{\mathcal{M}}$ and $D_{\mathcal{T}}$ generates $q_{\mathcal{M}}$, $k_{\mathcal{M}}$, $v_{\mathcal{M}}$, $q_{\mathcal{T}}$, $k_{\mathcal{T}}$, and $v_{\mathcal{T}}$. Then, the mutual blocks are operated:

$$\mathcal{M} = \mathcal{M} + \text{Softmax}(\frac{q_{\mathcal{M}} k_{\mathcal{T}}^{\top}}{\sqrt{d_k}}) v_{\mathcal{T}}, \tag{6.3}$$

$$\mathcal{T} = \mathcal{T} + \text{Softmax}(\frac{q_{\mathcal{T}} k_{\mathcal{M}}^{\top}}{\sqrt{d_k}}) v_{\mathcal{M}}. \tag{6.4}$$

The cross-attention layer is inserted into each self-attention-based DiT block to fuse the information from all modalities and achieve flexible training or inference.

6.2.4 Training Recipe

Step 1: Unconditional Pre-train. Since there are two independent DiTs in PackDiT, we can apply unconditional Pre-train on both of them to get better initial weights for the subsequent tasks. As shown in Algorithm 6.1, during unconditional pre-train, motion tokens and text tokens are sampled

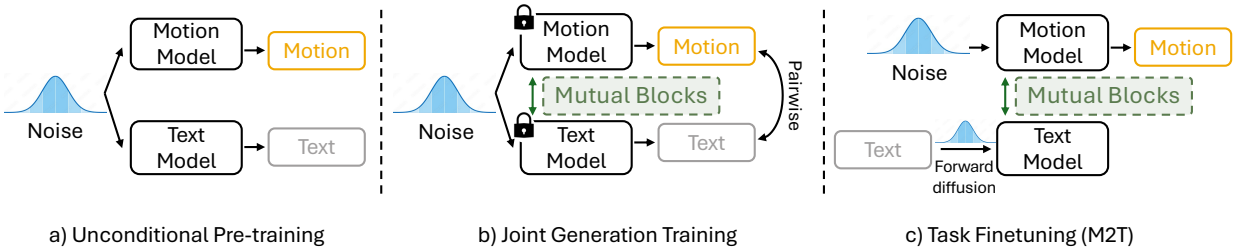


Figure 6.2: Training stages of the PackDiT model, illustrating the various phases, including a) unconditional pre-training, b) joint generation training, and c) task fine-tuning.

from the training dataset and are fed to $D_{\mathcal{M}}$ and $D_{\mathcal{T}}$ separately for standard unconditional diffusion training.

Step 2: Joint Generation Training. To make PackDiT able to conduct joint Motion and Text generation and better align the features from two modalities, we conduct joint generation training. We sample the same timestep $t_{\mathcal{M}} = t_{\mathcal{T}} \sim \text{Uniform}(1, \ldots, t)$ and the mutual blocks are enabled during training. Thus, both DiTs are trained together for feature alignment and joint generation.

Step 3: Motion-to-Text and Text-to-Motion Fine-tuning. As demonstrated in Algorithm 6.1, during training for either the Motion-to-Text or Text-to-Motion task, only the generating modality undergoes forward diffusion. In contrast, the condition modality is directly fed to the conditional DiT after Patchify, with the timestep set to 0. Consequently, through the mutual blocks between the two DiTs, PackDiT effectively performs reliable conditional generation.

Step 4: Joint Fine-Tuning. To train PackDiT on all four tasks, we employ a joint training approach, assigning a certain probability to each task at each iteration. For optimal performance in the evaluations detailed in Section 6.3, we further fine-tune the model on specific tasks, *i.e*, Text-to-Motion and Motion-to-Text generation, after the initial joint training phase. Therefore, the final training objective of PackDiT is to minimize the loss $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}_{\mathcal{M}} + \lambda \cdot \mathcal{L}_{\mathcal{T}}, \tag{6.5}$$

where λ is the term to balance the two objectives.

Table 6.2: Comparison of three motion-related tasks on HumanML3D [50] dataset. The evaluation metrics are computed using the encoders introduced in [50]. † indicates that LMM [212] is trained with additional data.

Methods	Source	Text-to-Motion			Motion Prediction		Motion In-between	
		R@1 (↑)	FID (↓)	DIV (↑)	FID (↓)	DIV (↑)	FID (↓)	DIV (↑)
Real Data	-	$0.511^{\pm.003}$	$0.002^{\pm.000}$	$9.503^{\pm.065}$	0.002	9.503	0.002	9.503
MLD [20]	CVPR'23	$0.481^{\pm.003}$	$0.473^{\pm.013}$	$\underline{9.724}^{\pm.082}$	-	-	-	-
T2M-GPT [207]	CVPR'23	$0.491^{\pm.003}$	$0.116^{\pm.004}$	$\mathbf{9.761}^{\pm.081}$	-	-	-	-
TM2T [51]	ECCV'22	$0.424^{\pm.017}$	$1.501^{\pm.003}$	$8.589^{\pm.076}$	-	-	-	-
MDM [174]	ICLR'23	$0.320^{\pm005}$	$0.544^{\pm.044}$	$9.559^{\pm.086}$	6.031	7.813	2.698	8.420
MotionGPT [63]	NeurIPS'23	$0.492^{\pm.003}$	$0.232^{\pm.008}$	$9.528^{\pm.071}$	0.905	8.972	0.214	**9.560**
LMM-Tiny† [212]	ECCV'24	$0.496^{\pm.002}$	$0.415^{\pm.002}$	$9.176^{\pm.074}$	-	-	-	-
LMM-Small† [212]	ECCV'24	$\underline{0.505}^{\pm.002}$	$\underline{0.227}^{\pm.002}$	$9.295^{\pm.076}$	-	-	-	-
TMT [133]	ECCV'24	$0.464^{\pm \text{ unk.}}$	$0.310^{\pm \text{ unk.}}$	$9.191^{\pm \text{ unk.}}$	-	-	-	-
PackDiT-Tiny	Ours	$0.498^{\pm.003}$	$0.232^{\pm.006}$	$9.381^{\pm.071}$	$\underline{0.764}$	**9.140**	$\underline{0.131}$	8.974
PackDiT-Small	Ours	$\mathbf{0.510}^{\pm.003}$	$\mathbf{0.106}^{\pm.006}$	$9.680^{\pm.078}$	**0.701**	$\underline{9.046}$	**0.119**	$\underline{9.114}$

6.3 Experimental Results

6.3.1 Evaluation Metrics and Datasets

Evaluation Metrics. Following [50, 174, 51, 63], Frechet Inception Distance (FID) is our primary metric for motion quality evaluation, which evaluates the feature distribution similarity between generated and real motions as detailed in [50]. Meanwhile, to measure the diversity of the generated motions, we use the Diversity (DIV) metric, which calculates the variance in features extracted from the motions as used in [50]. For text-motion retrieval evaluation, the accuracy of matching motions to their corresponding textual descriptions is assessed using the motion-retrieval precision (R Precision) metric, based on the feature space from [50], and measured by Top 1/2/3 retrieval accuracy. To evaluate the quality of generated motion captions, we adopt linguistic metrics

Table 6.3: Performance comparison on the **Motion-to-Text** task for the HumanML3D [50] dataset. All evaluation metrics are computed using encoders from [50].

Methods	R@3 (↑)	BLEU@4 (↑)	CIDEr (↑)
Real Data	0.828	-	-
LLM-based			
TM2T [51]	0.823	7.00	16.8
MotionGPT [63]	**0.827**	**12.47**	**29.2**
Diffusion-based			
PackDiT-Tiny	0.810	6.86	13.6
PackDiT-Small	<u>0.825</u>	<u>11.82</u>	<u>25.5</u>

from natural language processing studies as outlined in [51], including BLEU [125], and CIDEr [181].

Dataset. The HumanML3D dataset [50] is a comprehensive repository of 3D human motion sequences curated to advance research in human motion analysis and generation. It encompasses a diverse range of activities—including walking, running, dancing, and more complex actions—sourced from the AMASS dataset [110]. Comprising 14,616 motion sequences of durations between 2 and 10 seconds, each sequence is accompanied by multiple detailed textual annotations, enhancing its applicability for tasks such as text-to-motion and motion-to-text generation. The dataset incorporates a variety of actors to ensure broad representation across human movement patterns.

6.3.2 Training and Evaluation Details

We utilize a single NVIDIA A100 to train and evaluate PackDiT, which is developed on OpenDiT [216] and MotionGPT [63]. The number of parameters for each DiT of proposed PackDiT

Table 6.4: **Ablation study** on PackDiT. Evaluated PackDiT-Tiny on the HumanML3D dataset with the Text-to-Motion task. − means the hyperparameter or setup does not change compared with the baseline.

Experiments	Ablation Settings				Text-to-Motion	
	Dim_P	Text Encoder	Patch Size	w. Uncond	FID (↓)	R@1 (↑)
Baseline	64	BERT [29]	1	×	0.274	0.493
Projection Dim.	128	BERT [29]	1	×	0.264	0.493
	256	BERT [29]	1	×	0.251	0.495
Text Encoder	64	T5 [136]	1	×	0.618	0.472
Patch Size	64	BERT [29]	2	×	0.571	0.481
	64	BERT [29]	4	×	1.483	-
PackDiT-Tiny	64	BERT [29]	1	✓	**0.232**	**0.498**

Tiny and Small is around 75M and 120M, which are similar to the setup of LMM [212]. With the batch size as 128, PackDiT is trained with the Adam [72] optimizer, and the learning rate is set to 10^{-4}. The patch size is set to 1 for both $D_{\mathcal{M}}$ and $D_{\mathcal{T}}$ during evaluation, and more patch size setup is discussed in Ablation Study 6.4.1. As mentioned in Sec 5.2, a text encoder, a text decoder, and a projection model are used in the Text Generation pipeline. Following [4], BERT [29], and GPT-2 [135] are used as the text encoder and text decoder, respectively. The projection model, P, is trained with projection dimension 64 when the encoder and decoder models are frozen. The unconditional pre-train takes around 10 epochs. Then, the PackDiT is jointly trained with all tasks for 200 epochs. To achieve the best performance of PackDiT, the models used for evaluation are fine-tuned on specific tasks for 300 epochs after joint training with all tasks. The motion representations of PackDiT follows [50, 63] for fair comparison.

6.3.3 Experimental Results of All Tasks

Evaluation on Motion-related Tasks. Our proposed method is compared with other SOTA methods on the HumanML3D [50] dataset. As shown in Table 6.2, our PackDiT-Tiny and PackDiT-Small achieve 0.498 and 0.504 R@1 on text-to-motion task, which demonstrates a comparable performance with previous SOTA method LMM [212], while using a smaller amount of training data. In other tasks like motion prediction and motion in-between tasks, PackDiT achieves 0.701 and 0.119 FID scores, outperforming MotionGPT [63] by a large margin.

Evaluation on Motion-to-Text. We also evaluate PackDiT's performance on the motion-to-text task. Despite utilizing a diffusion-based backbone, our proposed method demonstrates effective motion-to-text generation capabilities. PackDiT-small achieves an R@3 of 0.784, a BLEU@4 score of 8.12, and a CIDEr score of 15.5. These results are comparable to those of LLM-based methods, highlighting PackDiT's competitive performance despite the inherent challenges of using a diffusion model for language tasks.

Visualization Results. As shown in Fig 6.3, we present additional Text-to-Motion generation results. These results demonstrate that our method, PackDiT, is capable of generating diverse and reliable motion sequences. The generated motions exhibit a wide variety of actions and behaviors, accurately reflecting the nuances of the input text descriptions. The visualization videos are attached to the appendix.

6.4 More Qualitative Results

Motion-to-Text visualization results are presented in Fig. 6.2, demonstrating the effectiveness of our proposed method. Similarly, extended Text-to-Motion visualization results are shown in Fig. 6.5, highlighting the ability of PackDiT to generate diverse and temporally stable motions that adhere closely to the given descriptions. Visualization results for the Motion In-Between task are provided in Fig. 6.3, emphasizing PackDiT's capability to produce smooth and contextually coherent intermediate motions. Moreover, Fig. 6.4 showcases the results of the Motion Prediction

task, illustrating the model's ability to accurately predict plausible future motions based on prior sequences. These results demonstrate that PackDiT achieves high-quality performance across four tasks, including Motion-to-Text, Text-to-Motion, Motion In-Between, and Motion Prediction, underscoring its stability and robustness.

6.4.1 Ablation Studies

To further analyze the PackDiT, we conduct ablation study on several hyperparameters, training strategies and alternative models used by us. All ablation studies are evaluated on Text-to-Motion tasks with PackDiT-Tiny as default.

Target Dimension of Projection Model. As shown in Tab 6.4, the best performance is achieved when the target dimension of the projection model, P, is set to 256. 128 and 256 are closer to the dimension of BERT's hidden states and save more information from the original text tokens. However, based on the trade-off between accuracy and efficiency, we choose $\text{Dim}_P = 64$ when we are training PackDiT-Tiny.

Text Encoder. Following [4], we utilize BERT [29] as our text encoder for the text generation pipeline. To further investigate the PackDiT, T5 [136], an Encoder-Decoder based Large Language Model, is applied to our text pipeline for a comparison. Only the encoder part of the T5-base is integrated. As illustrated in Tab 6.4, the performance of Text-to-Motion with BERT surpasses the T5 version by a remarkable margin since the pre-train weights of T5 are based on translation, summarization, question answering, and *etc*, which may not be suitable for motion-related tasks.

Patch Size. We change the patch size of motion diffusion model and conduct an ablation study. According to Tab 6.4, the patch size 1 provides the best performance while patch size 4 significantly impacts the performance. We find that once the dimension of input tokens is similar to the hidden dimension of DiTs, the generation results are not reliable.

Table 6.5: The comparison of the number of parameters. The PackDiT includes both text and motion diffusion models.

Methods	Arch.	# Param.
MotionGPT [63]	AR	248M
LMM-Tiny [212]	Diffusion	90M
LMM-Small [212]	Diffusion	160M
PackDiT-Tiny	Diffusion	72M
PackDiT-Small	Diffusion	229M

Unconditional Pre-train. As indicated in Table 6.4, the Unconditional Pre-train improves the final Text-to-Motion Generation performance. This pre-training phase allows PackDiT to develop a better understanding of the underlying data distribution, which is crucial for generating realistic and coherent motions. By initializing the model with weights that are already attuned to the data characteristics, the subsequent training process is more efficient and effective. We plan to utilize more unpaired motion sequences and motion descriptions for future works and further improve the effectiveness of the Unconditional Pre-train.

Number of Parameters To make a fair comparison with other SOTA motion generation models, we compare the number of parameters of PackDiT with other methods. As shown in Tab 6.5, we compare the number of parameters with MotionGPT [63] and LMM [212], which shows a fair comparison with other SOTA methods.

6.5 Limitations

The performance of PackDiT heavily relies on the quality and diversity of the HumanML3D dataset. Limited data variety may hinder the generalization of the model. Also, current evaluation metrics such as FID and Recall may not fully capture the quality and realism of generated motions and texts, suggesting a need for more comprehensive performance assessment methods.

6.6 Conclusion

In this work, we presented **PackDiT**, a novel diffusion-based framework for joint human motion and text generation. PackDiT's unique dual Diffusion Transformer (DiT) architecture with mutual blocks enables efficient handling of multiple generation tasks, including text-to-motion, motion-to-text, motion prediction, and motion in-between generation within a unified model structure. This approach addresses critical limitations in previous models, which often restricted generation to single modalities or required complex configurations. Extensive experiments on the HumanML3D dataset demonstrate that PackDiT achieves state-of-the-art results, particularly in text-to-motion generation with an FID score of 0.106, as well as strong performance in motion prediction and in-between tasks. The mutual prompting mechanism facilitates enhanced information exchange, enabling PackDiT to produce high-quality, diverse outputs that closely align with human-generated data. PackDiT thus establishes a flexible and robust foundation for multi-modal generation, with broad implications for synthetic data creation, immersive environments, and human-computer interaction applications, setting a new standard for joint motion and text synthesis.

Algorithm 6.1: The **pseudo-code** of different training stages of PackDiT depends on different tasks, *e.g*, unconditional pre-train, Text-to-Motion, and Motion-to-Text.

Algorithm 5 Unconditional Pre-train

Require: Motion tokens, $\mathcal{M}$, Text tokens, $\mathcal{T}$

Repeat

$t_{\mathcal{M}}, t_{\mathcal{T}} \sim \text{Uniform}(1, \dots, t)$

$\epsilon_{\mathcal{M}}, \epsilon_{\mathcal{T}} \sim \mathcal{N}(0, 1)$

$\mathcal{M}_{t_{\mathcal{M}}} = \sqrt{\overline{\alpha}_{t_{\mathcal{M}}}}\mathcal{M} + \sqrt{1 - \overline{\alpha}_{t_{\mathcal{M}}}}\epsilon_{\mathcal{M}}$

$\mathcal{T}_{t_{\mathcal{T}}} = \sqrt{\overline{\alpha}_{t_{\mathcal{T}}}}\mathcal{T} + \sqrt{1 - \overline{\alpha}_{t_{\mathcal{T}}}}\epsilon_{\mathcal{T}}$

Take gradient step on

$\nabla_{D_{\mathcal{M}}} \|\epsilon - D_{\mathcal{M}}(\mathcal{M}_{t_{\mathcal{M}}}, t_{\mathcal{M}})\|_2^2$

$\nabla_{D_{\mathcal{T}}} \|\epsilon - D_{\mathcal{T}}(\mathcal{T}_{t_{\mathcal{T}}}, t_{\mathcal{T}})\|_2^2$

until converge

Algorithm 6 Joint Generation Training

Require: Paired Motion tokens, $\mathcal{M}$, Text tokens, $\mathcal{T}$

Repeat

$t_{\mathcal{M}} = t_{\mathcal{T}} \sim \text{Uniform}(1, \dots, t)$

$\epsilon_{\mathcal{M}}, \epsilon_{\mathcal{T}} \sim \mathcal{N}(0, 1)$

$\mathcal{M}_{t_{\mathcal{M}}} = \sqrt{\overline{\alpha}_{t_{\mathcal{M}}}}\mathcal{M} + \sqrt{1 - \overline{\alpha}_{t_{\mathcal{M}}}}\epsilon_{\mathcal{M}}$

$\mathcal{T}_{t_{\mathcal{T}}} = \sqrt{\overline{\alpha}_{t_{\mathcal{T}}}}\mathcal{T} + \sqrt{1 - \overline{\alpha}_{t_{\mathcal{T}}}}\epsilon_{\mathcal{T}}$

Take gradient step on

$\nabla_{D_{\mathcal{M}}} \|\epsilon - D_{\mathcal{M}}(\mathcal{M}_{t_{\mathcal{M}}}, t_{\mathcal{M}}), D_{\mathcal{T}}(\mathcal{T}_{t_{\mathcal{T}}}, t_{\mathcal{T}}))\|_2^2$

$\nabla_{D_{\mathcal{T}}} \|\epsilon - D_{\mathcal{T}}(\mathcal{T}_{t_{\mathcal{T}}}, t_{\mathcal{T}}, D_{\mathcal{M}}(\mathcal{M}_{t_{\mathcal{M}}}, t_{\mathcal{M}}))\|_2^2$

until converge

Algorithm 7 Text-to-Motion Training

Require: Paired Motion tokens, $\mathcal{M}$, Text tokens, $\mathcal{T}$

Repeat

$t_{\mathcal{M}} \sim \text{Uniform}(1, \dots, t)$

$\epsilon_{\mathcal{M}} \sim \mathcal{N}(0, 1)$

$\mathcal{M}_{t_{\mathcal{M}}} = \sqrt{\overline{\alpha}_{t_{\mathcal{M}}}}\mathcal{M} + \sqrt{1 - \overline{\alpha}_{t_{\mathcal{M}}}}\epsilon_{\mathcal{M}}$

Take gradient step on

$\nabla_{D_{\mathcal{M}}} \|\epsilon - D_{\mathcal{M}}(\mathcal{M}_{t_{\mathcal{M}}}, t_{\mathcal{M}}, D_{\mathcal{T}}(\mathcal{T}, 0)\|_2^2$

until converge

Algorithm 8 Joint Training

Require: Motion tokens, $\mathcal{M}$, Text tokens, $\mathcal{T}$

Repeat

task = RandomChoise(t2m, m2t, uncond, joint)

$t_{\mathcal{M}}, t_{\mathcal{T}} \sim \text{Uniform}(1, \dots, t)$

$\epsilon_{\mathcal{M}}, \epsilon_{\mathcal{T}} \sim \mathcal{N}(0, 1)$

Train$(D_{\mathcal{M}}, D_{\mathcal{T}}, \mathcal{M}, \mathcal{T}, t_{\mathcal{M}}, t_{\mathcal{T}}, \epsilon_{\mathcal{M}}, \epsilon_{\mathcal{T}}, \text{task})$

until converge

Motion to Text			
			
GT	a person is **walking** at an angle to the **right**.	a person is **bent forward** with arms dangling in front of them	a person is cheering and **dancing**.
Ours	a person **walks** forward and turn slightly to their **right**.	a person **bends** over **forward** and picks something up loosely	a person does a workout **dance**.
			
GT	the person was **pushed** but stayed standing.	a person **sits down** in a hurry.	a **walking** person suddenly gets staggered to their **right**, then recovers.
Ours	the person is **pushed** back.	a person is **sitting down** a chair.	a person is **walking** forward straight and then slumped to the **right**.

Figure 6.2: More Motion-to-Text visualization results of PackDiT on HumanML3D dataset.

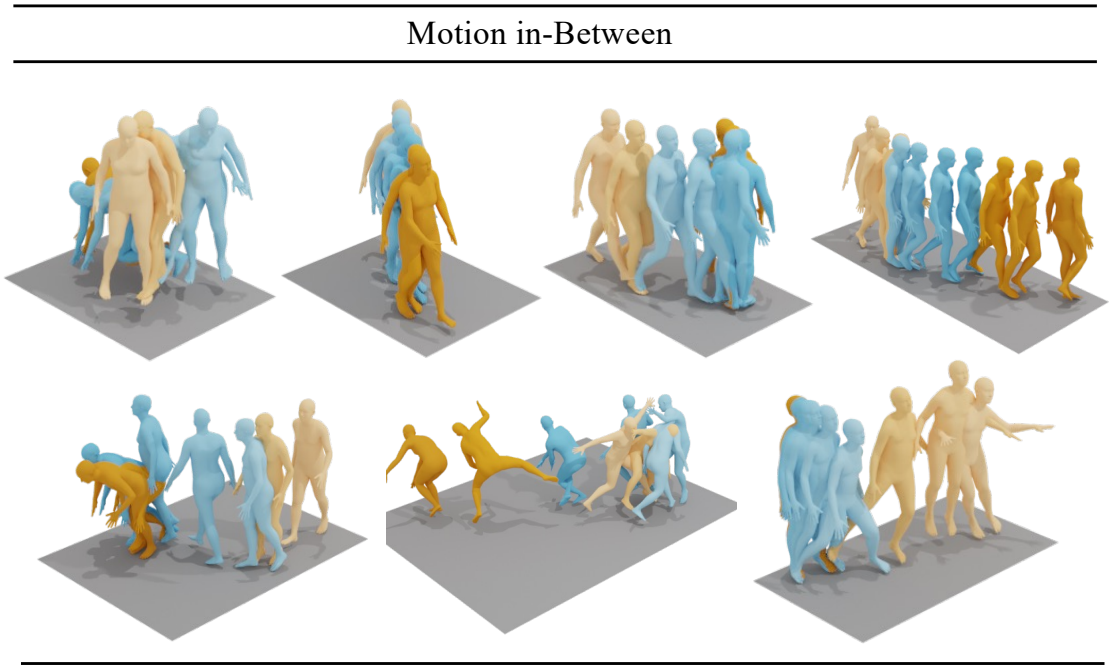


Figure 6.3: More Motion in-Between visualization results of PackDiT. The orange avatars are from the ground truth motion, while the blue ones are generated by PackDiT.

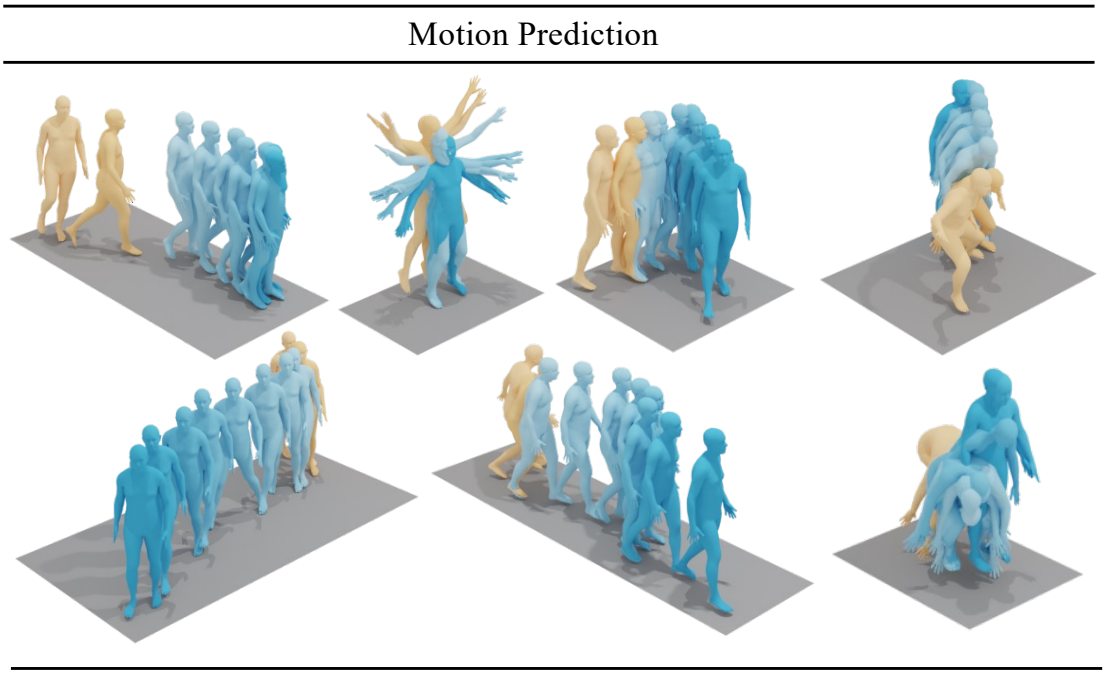


Figure 6.4: More Motion Prediction visualization results of PackDiT. The orange avatars are from the ground truth motion, while the blue ones are generated by PackDiT.

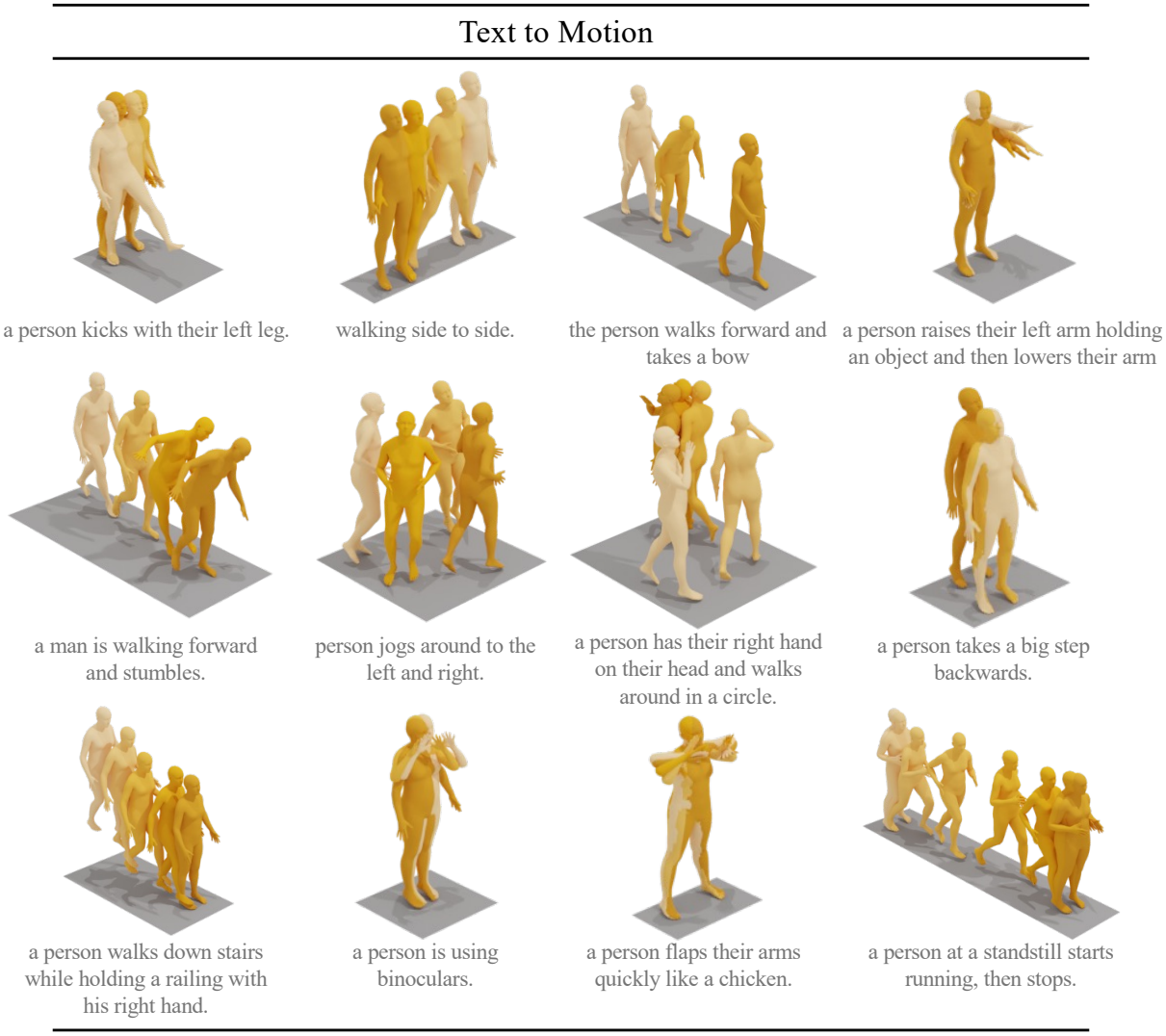


Figure 6.5: More Text-to-Motion visualization results of PackDiT.

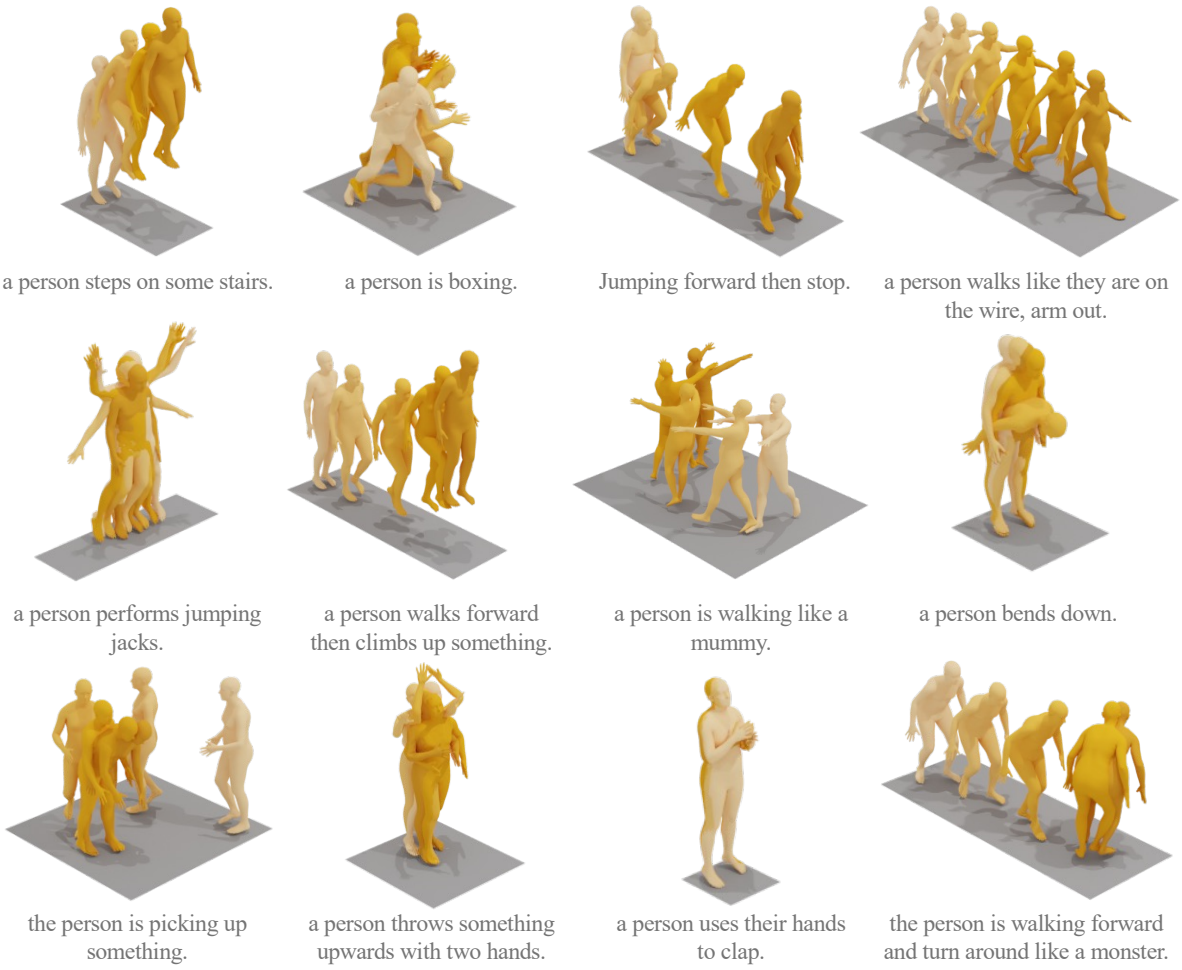


Figure 6.3: Visualization results of Text-to-Motion Generation via PackDiT.

Chapter 7

CONCLUSION

This dissertation presents several key contributions toward advancing human pose estimation (HPE) and motion generation. First, we introduce GolfPose, a lightweight temporal-based 2D HPE pipeline designed for golf swing analysis. Unlike conventional 2D HPE methods, which struggle with motion blur and severe self-occlusion, GolfPose leverages temporal information to improve keypoint estimation accuracy while maintaining computational efficiency. The proposed model is lightweight and optimized for edge-device deployment, enabling real-time golf swing analysis.

Second, we propose ZeDO, a zero-shot diffusion-based optimization framework for 3D HPE that eliminates domain gaps by leveraging a pre-trained diffusion pose generation model. Unlike learning-based methods, which suffer from performance degradation in cross-domain settings, ZeDO iteratively refines 3D pose estimates by minimizing 2D re-projection errors, ensuring robust generalization without requiring direct 2D-3D supervision. The experimental results demonstrate state-of-the-art performance on multiple benchmark datasets, highlighting its effectiveness in domain-agnostic 3D pose estimation.

Third, we develop UniHPR, a unified pose representation learning framework that aligns embeddings from 2D human poses, 3D human poses, and images within a shared feature space. By leveraging contrastive learning, UniHPR enables a modality-agnostic HPE pipeline that seamlessly integrates both 2D and 3D pose estimation tasks. The proposed singular value-based contrastive learning strategy enhances feature alignment, improving robustness across diverse datasets and domains.

Finally, we introduce PackDiT, a diffusion-based framework for joint motion and text generation, designed to synchronize motion synthesis with textual descriptions. Unlike existing methods that operate in a unidirectional manner (either text-to-motion or motion-to-text), PackDiT employs

mutually conditioned Diffusion Transformers (DiTs) to enable bidirectional motion-text generation. By utilizing a shared latent space, the model achieves state-of-the-art performance on HumanML3D, setting a new benchmark for multi-modal motion understanding and synthesis.

Collectively, these contributions push the boundaries of robust, efficient, and generalizable human pose estimation while extending its impact to motion synthesis and cross-modal learning. Through the integration of optimization techniques, contrastive learning, and diffusion-based generative models, this work lays the foundation for future advancements in human-centric AI, with applications spanning sports analysis, robotics, AR/VR, and human-computer interaction.

BIBLIOGRAPHY

[1] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5915–5920. IEEE, 2018. 8

[2] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5167–5176, 2018. 1, 40

[3] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3395–3404, 2019. 53

[4] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, pages 1692–1717. PMLR, 2023. 7, 58, 59, 63, 65

[5] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, pages 1692–1717. PMLR, 2023.

[6] Rajendra Bhatia. Matrix analysis, 1997.

[7] 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. 5, 21, 22, 25, 26, 36, 40

[8] Lewis Bridgeman, Marco Volino, Jean-Yves Guillemaut, and Adrian Hilton. Multi-person 3d pose estimation and tracking in sports. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 9, 21

[9] Jiarui Cai, Yizhou Wang, Hung-Min Hsu, Haotian Zhang, and Jenq-Neng Hwang. Dior: Distill observations to representations for multi-object tracking and segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 520–529, 2022. 12

[10] Jiarui Cai, Mingze Xu, Wei Li, Yuanjun Xiong, Wei Xia, Zhuowen Tu, and Stefano Soatto. Memot: Multi-object tracking with memory. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8090–8100, 2022. 6

[11] Shidong Cao, Wenhao Chai, Shengyu Hao, and Gaoang Wang. Image reference-guided fashion design with structure-aware transfer by diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3524–3528, 2023. 7

[12] Shidong Cao, Wenhao Chai, Shengyu Hao, Yanting Zhang, Hangyue Chen, and Gaoang Wang. Difffashion: Reference-based fashion design with structure-aware transfer by diffusion models. *arXiv preprint arXiv:2302.06826*, 2023. 7

[13] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 3, 4

[14] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 3, 4

[15] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 6

[16] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23040–23050, 2023. 7

[17] Wenhao Chai, Zhongyu Jiang, Jenq-Neng Hwang, and Gaoang Wang. Global adaptation meets local generalization: Unsupervised domain adaptation for 3d human pose estimation. *arXiv preprint arXiv:2303.16456*, 2023. 8, 22, 29, 30, 35, 37, 40, 47

[18] 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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5707–5717, 2019.

[19] Haoming Chen, Runyang Feng, Sifan Wu, Hao Xu, Fengcheng Zhou, and Zhenguang Liu. 2d human pose estimation: A survey. *Multimedia Systems*, 29(5):3115–3138, 2023. 3

[20] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 56, 61

[21] 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

[22] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5386–5395, 2020. 4, 5

[23] Vasileios Choutas, Federica Bogo, Jingjing Shen, and Julien Valentin. Learning to fit morphable models. In *European Conference on Computer Vision*, pages 160–179. Springer, 2022. 5, 35

[24] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1831–1840, 2017. 3

[25] Hai Ci, Xiaoxuan Ma, Chunyu Wang, and Yizhou Wang. Locally connected network for monocular 3d human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1429–1442, 2020. 21

[26] Hai Ci, Mingdong Wu, Wentao Zhu, Xiaoxuan Ma, Hao Dong, Fangwei Zhong, and Yizhou Wang. Gfpose: Learning 3d human pose prior with gradient fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4800–4810, 2023. 21, 23, 25, 27, 28, 30, 36, 37

[27] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020.

[28] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9760–9770, 2023. 8

[29] 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. 58, 63, 65

[30] Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3d human pose estimation: motion to the rescue. *Advances in Neural Information Processing Systems*, 32, 2019. 53

[31] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4, 6

[32] Haodong Duan, Jiaqi Wang, Kai Chen, and Dahua Lin. Pyskl: Towards good practices for skeleton action recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7351–7354, 2022. 1, 40

[33] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2969–2978, 2022.

[34] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14676–14686, 2021. 4

[35] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14676–14686, 2021.

[36] Zigang Geng, Chunyu Wang, Yixuan Wei, Ze Liu, Houqiang Li, and Han Hu. Human pose as compositional tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 660–671, 2023.

[37] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Košecká, and Ziyan Wu. Hierarchical kinematic human mesh recovery. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 768–784. Springer, 2020. 40, 53

[38] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Košecká, and Ziyan Wu. Hierarchical kinematic human mesh recovery. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 768–784. Springer, 2020.

[39] Mohsen Gholami, Bastian Wandt, Helge Rhodin, Rabab Ward, and Z Jane Wang. Adaptpose: Cross-dataset adaptation for 3d human pose estimation by learnable motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13075–13085, 2022. 21, 22, 30, 35, 37, 47

[40] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 7

[41] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023. 6

[42] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13041–13051, 2023. 21, 23, 27, 36

[43] 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. 21, 22, 29, 30, 35, 36, 37

[44] 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. 40, 47

[45] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 7

[46] Renshu Gu. *Towards Multi-Person 3D Pose Estimation in Natural Videos*. University of Washington, 2020. 26, 36

[47] Renshu Gu, Zhongyu Jiang, Gaoang Wang, Kevin McQuade, and Jenq-Neng Hwang. Unsupervised universal hierarchical multi-person 3d pose estimation for natural scenes. *Multimedia Tools and Applications*, 81(23):32883–32906, 2022. 22, 25

[48] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10884–10894, 2019. 6, 53

[49] Chuan Guo, Yuxuan Mu, Xinxin Zuo, Peng Dai, Youliang Yan, Juwei Lu, and Li Cheng. Generative human motion stylization in latent space. *arXiv preprint arXiv:2401.13505*, 2024. 8

[50] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, June 2022. 55, 56, 57, 61, 62, 63, 64

[51] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pages 580–597. Springer, 2022. 55, 56, 61, 62

[52] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. 55

[53] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022. 7

[54] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[55] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 7, 55

[56] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020. 25, 27, 31

[57] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 44

[58] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[59] Nhat M Hoang, Kehong Gong, Chuan Guo, and Michael Bi Mi. Motionmix: Weakly-supervised diffusion for controllable motion generation. *arXiv preprint arXiv:2401.11115*, 2024. 8

[60] Jun Hu, Zhongyu Jiang, Xionghao Ding, Taijiang Mu, and Peter Hall. Vgpn: Voice-guided pointing robot navigation for humans. In *2018 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1107–1112. IEEE, 2018. 1, 40

[61] 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. 28, 29, 30, 41, 48, 49

[62] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7718–7727, 2019. 28

[63] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36, 2024. 8, 55, 56, 57, 61, 62, 63, 64, 66

[64] Zhongyu Jiang, Haorui Ji, Samuel Menaker, and Jenq-Neng Hwang. Golfpose: Golf swing analyses with a monocular camera based human pose estimation. In *2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2022. 1, 21, 40

[65] Zhongyu Jiang, Haorui Ji, Cheng-Yen Yang, and Jenq-Neng Hwang. 2d human pose estimation calibration and keypoint visibility classification. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6095–6099. IEEE, 2024. 40

[66] Zhongyu Jiang, Zhuoran Zhou, Lei Li, Wenhao Chai, Cheng-Yen Yang, and Jenq-Neng Hwang. Back to optimization: Diffusion-based zero-shot 3d human pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6142–6152, 2024. 8

[67] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *bmvc*, volume 2, page 5. Aberystwyth, UK, 2010. 30

[68] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *2021 International Conference on 3D Vision (3DV)*, pages 42–52. IEEE, 2021.

[69] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 37, 53

[70] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5614–5623, 2019. 53

[71] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4):1–14, 2023.

[72] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 63

[73] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 7

[74] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020. 21, 35, 53

[75] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11127–11137, 2021. 30, 35, 53

[76] 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. 6, 21, 30, 35

[77] 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 *ICCV*, 2019. 40, 53

[78] 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 (ICCV)*, October 2019.

[79] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 53

[80] Jogendra Nath Kundu, Siddharth Seth, Varun Jampani, Mugalodi Rakesh, R Venkatesh Babu, and Anirban Chakraborty. Self-supervised 3d human pose estimation via part guided novel image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6152–6162, 2020.

[81] Jogendra Nath Kundu, Siddharth Seth, Mayur Rahul, M. Rakesh, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Kinematic-structure-preserved representation for unsupervised 3d human pose estimation. *ArXiv*, abs/2006.14107, 2020.

[82] Jin Han Lee, Sehyung Lee, Guoxuan Zhang, Jongwoo Lim, Wan Kyun Chung, and Il Hong Suh. Outdoor place recognition in urban environments using straight lines. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5550–5557. IEEE, 2014. 14

[83] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9887–9895, 2019.

[84] Jiefeng Li, Siyuan Bian, Chao Xu, Zhicun Chen, Lixin Yang, and Cewu Lu. Hybrik-x: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery. *arXiv preprint arXiv:2304.05690*, 2023. 1, 40

[85] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3383–3393, 2021. 30, 35

[86] 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. 5, 29

[87] Sijin Li, Weichen Zhang, and Antoni B Chan. Maximum-margin structured learning with deep networks for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2848–2856, 2015. 36

[88] 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*, 25:1282–1293, 2022. 47

[89] 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*, 25:1282–1293, 2022.

[90] 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. 37

[91] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, pages 590–606. Springer, 2022. 6, 21, 36, 40, 53

[92] Junfan Lin, Jianlong Chang, Lingbo Liu, Guanbin Li, Liang Lin, Qi Tian, and Chang-wen Chen. Being comes from not-being: Open-vocabulary text-to-motion generation with wordless training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23222–23231, 2023.

[93] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12939–12948, 2021.

[94] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[95] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 15

[96] 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.

[97] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *European Conference on Computer Vision*, pages 388–404. Springer, 2022. 7

[98] Hanbing Liu, Jun-Yan He, Zhi-Qi Cheng, Wangmeng Xiang, Qize Yang, Wenhao Chai, Gaoang Wang, Xu Bao, Bin Luo, Yifeng Geng, et al. Posynda: Multi-hypothesis pose synthesis domain adaptation for robust 3d human pose estimation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5542–5551, 2023. 8

[99] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6

[100] Zhenguang Liu, Haoming Chen, Runyang Feng, Shuang Wu, Shouling Ji, Bailin Yang, and Xun Wang. Deep dual consecutive network for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 525–534, 2021.

[101] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 5, 21

[102] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015. 8

[103] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

[104] Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Humantomato: Text-aligned whole-body motion generation. *arXiv preprint arXiv:2310.12978*, 2023.

[105] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022.

[106] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. volume 508, pages 293–304. Elsevier, 2022. 7

[107] Zhengxiong Luo, Zhicheng Wang, Yan Huang, Liang Wang, Tieniu Tan, and Erjin Zhou. Rethinking the heatmap regression for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13264–13273, 2021.

[108] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5137–5146, 2018.

[109] Xiaoxuan Ma, Jiajun Su, Chunyu Wang, Wentao Zhu, and Yizhou Wang. 3d human mesh estimation from virtual markers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 534–543, 2023. 35

[110] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 55, 62

[111] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, October 2019.

[112] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017. 5, 36, 37, 40, 47

[113] 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 *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 28, 29, 30, 41, 48, 49

[114] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *Acm transactions on graphics (tog)*, 36(4):1–14, 2017. 37

[115] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8844–8854, 2022. 6

[116] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. 27

[117] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[118] Lea Muller, Ahmed AA Osman, Siyu Tang, Chun-Hao P Huang, and Michael J Black. On self-contact and human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9990–9999, 2021. 5, 22, 26, 37

[119] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *Advances in neural information processing systems*, 30, 2017. 4

[120] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 483–499. Springer, 2016. 3

[121] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 38

[122] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. 2018. 53

[123] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[124] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *Proceedings of the European conference on computer vision (ECCV)*, pages 269–286, 2018. 4

[125] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 62

[126] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 48

[127] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7025–7034, 2017.

[128] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 459–468, 2018. 53

[129] 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. 5, 21, 29, 36, 37, 40, 47

[130] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 7, 58

[131] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 21

[132] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[133] Yijun Qian, Jack Urbanek, Alexander Hauptmann, and Jungdam Won. Text motion translator: A bi-directional model for enhanced 3d human motion generation from open-vocabulary descriptions. In *European Conference on Computer Vision*, 2024. 61

[134] 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. 7, 41

[135] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 58, 63

[136] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 63, 65

[137] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[138] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015. 7

[139] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. Learning monocular 3d human pose estimation from multi-view images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8437–8446, 2018. 9, 28

[140] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[141] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 7

[142] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10219–10228, 2023. 59

[143] Saeed Saadatnejad, Ali Rasekh, Mohammadreza Mofayezi, Yasamin Medghalchi, Sara Rajabzadeh, Taylor Mordan, and Alexandre Alahi. A generic diffusion-based approach for 3d human pose prediction in the wild. *arXiv preprint arXiv:2210.05669*, 2022.

[144] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 1, 40

[145] Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. Monocular 3d human pose estimation by generation and ordinal ranking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2325–2334, 2019. 28

[146] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017.

[147] Michael Snower, Asim Kadav, Farley Lai, and Hans Peter Graf. 15 keypoints is all you need. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6738–6748, 2020. 1, 40

[148] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 7

[149] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265, 2015.

[150] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 7, 55, 58

[151] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 25, 27, 31, 38

[152] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 44

[153] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *European Conference on Computer Vision*, pages 744–760. Springer, 2020. 5, 35, 36

[154] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 7, 55

[155] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 24

[156] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 44

[157] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 7, 24, 27, 29, 31, 38, 47, 55

[158] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 89–98, 2018.

[159] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 3, 16, 40

[160] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, 2021. 6

[161] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, One-stage, Regression of Multiple 3D People. In *ICCV*, 2021. 40

[162] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13243–13252, 2022. 6, 21

[163] 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 *CVPR*, 2022. 40

[164] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5349–5358, 2019. 53

[165] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*, 2016. 59

[166] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 581–600. Springer, 2020.

[167] Jun Kai Vince Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3d human body shape and pose prediction. 2017. 6

[168] Zheng Tang, Renshu Gu, and Jenq-Neng Hwang. Joint multi-view people tracking and pose estimation for 3d scene reconstruction. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018. 5, 22, 25, 26

[169] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36, 2024. 59

[170] Bugra Tekin, Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 3941–3950, 2017.

[171] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022. 40

[172] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 40

[173] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[174] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 55, 56, 61

[175] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014. 3

[176] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 448–458, 2023. 55

[177] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. *Advances in Neural Information Processing Systems*, 30, 2017. 6

[178] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[179] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[180] 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. 6, 42, 43

[181] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 62

[182] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018. 28, 29, 49

[183] Bastian Wandt, James J Little, and Helge Rhodin. Elepose: Unsupervised 3d human pose estimation by predicting camera elevation and learning normalizing flows on 2d poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6635–6645, 2022.

[184] Bastian Wandt, Marco Rudolph, Petrissa Zell, Helge Rhodin, and Bodo Rosenhahn. Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13294–13304, 2021.

[185] Chunyu Wang, Yizhou Wang, Zhouchen Lin, Alan L Yuille, and Wen Gao. Robust estimation of 3d human poses from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2361–2368, 2014. 36

[186] 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.

[187] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 3

[188] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 41, 43

[189] Kang Wang, Rui Zhao, and Qiang Ji. Human computer interaction with head pose, eye gaze and body gestures. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 789–789. IEEE, 2018. 1, 40

[190] Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Probabilistic monocular 3d human pose estimation with normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11199–11208, 2021. 28

[191] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016. 3

[192] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 16210–16220, 2022. 1, 21, 40

[193] Scott Workman, Connor Greenwell, Menghua Zhai, Ryan Baltenberger, and Nathan Jacobs. Deepfocal: A method for direct focal length estimation. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 1369–1373. IEEE, 2015.

[194] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. 3

[195] Tianhan Xu and Wataru Takano. Graph stacked hourglass networks for 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16105–16114, 2021.

[196] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7760–7770, 2019. 53

[197] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 4, 6

[198] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1179–1189, 2023.

[199] 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*, pages 2924–2933, 2023. 5, 21

[200] Zhenbo Yu, Bingbing Ni, Jingwei Xu, Junjie Wang, Chenglong Zhao, and Wenjun Zhang. Towards alleviating the modeling ambiguity of unsupervised monocular 3d human pose estimation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8631–8631, 2021.

[201] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution vision transformer for dense predict. *Advances in Neural Information Processing Systems*, 34:7281–7293, 2021. 4

[202] Ailing Zeng, Xiao Sun, Fuyang Huang, Minhao Liu, Qiang Xu, and Stephen Lin. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 507–523. Springer, 2020.

[203] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang. 3d human mesh regression with dense correspondence. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7054–7063, 2020. 53

[204] Yu Zhan, Fenghai Li, Renliang Weng, and Wongun Choi. Ray3d: ray-based 3d human pose estimation for monocular absolute 3d localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13116–13125, 2022. 22

[205] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 40, 53

[206] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11446–11456, 2021. 40, 53

[207] 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. *arXiv preprint arXiv:2301.06052*, 2023. 55, 56, 61

[208] 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. 37, 47

[209] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[210] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.

[211] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 8, 40, 55, 56

[212] Mingyuan Zhang, Daisheng Jin, Chenyang Gu, Fangzhou Hong, Zhongang Cai, Jingfang Huang, Chongzhi Zhang, Xinying Guo, Lei Yang, Ying He, et al. Large motion model for unified multi-modal motion generation. *arXiv preprint arXiv:2404.01284*, 2024. 8, 55, 56, 61, 63, 64, 66

[213] Zhe Zhang, Jie Tang, and Gangshan Wu. Simple and lightweight human pose estimation. *arXiv preprint arXiv:1911.10346*, 2019. 13, 16

[214] Zhenyu Zhang, Wenhao Chai, Zhongyu Jiang, Tian Ye, Mingli Song, Jenq-Neng Hwang, and Gaoang Wang. Mpm: A unified 2d-3d human pose representation via masked pose modeling. *arXiv preprint arXiv:2306.17201*, 2023. 7, 47

[215] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3425–3435, 2019. 5, 21, 29, 36, 40, 47

[216] Xuanlei Zhao, Zhongkai Zhao, Ziming Liu, Haotian Zhou, Qianli Ma, and Yang You. Open-dit: An easy, fast and memory-efficient system for dit training and inference. https://github.com/NUS-HPC-AI-Lab/OpenDiT, 2024. 62

[217] Zhonghan Zhao, Wenhao Chai, Shengyu Hao, Wenhao Hu, Guanhong Wang, Shidong Cao, Mingli Song, Jenq-Neng Hwang, and Gaoang Wang. A survey of deep learning in sports applications: Perception, comprehension, and decision. *arXiv preprint arXiv:2307.03353*, 2023. 1, 40

[218] 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. 6

[219] 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. 21

[220] 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. 40

[221] Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Truncated diffusion probabilistic models and diffusion-based adversarial auto-encoders. *arXiv preprint arXiv:2202.09671*, 2022. 27

[222] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C Karen Liu, and Leonidas J Guibas. Gimo: Gaze-informed human motion prediction in context. In *European Conference on Computer Vision*, pages 676–694. Springer, 2022.

[223] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2344–2353, 2019.

[224] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 4, 5

[225] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15085–15099, 2023. 6

[226] Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 55

[227] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 6