Learning Knowledge from Textual Descriptions for 3D Human Pose Estimation

Yi Wu [1], Jingtian Li [1], Shangfei Wang [1*], Guoming Li [2], Meng Mao [2], Linxiang Tan [2]

[1]School of Computer Science and Technology, University of Science and Technology of China
[2]China Merchants Bank
{wy221711, lijingtian}@mail.ustc.edu.cn, sfwang@ustc.edu.cn
{lkm, melvinmaonn, tanlinxiang252}@cmbchina.com

Abstract

Mainstream 3D human pose estimation methods directly predict 3D coordinates of joints from 2D keypoints, suffering from severe depth ambiguity. Pose textual descriptions contain abundant semantic information, which facilitates the model to learn the spatial relationship among different body parts, partially alleviating this issue. Leveraging this insight, we propose a 3D human pose estimation method assisted by textual descriptions. Specifically, we utilize an automatic captioning pipeline to generate textual descriptions of 3D poses based on spatial relations among joints. These descriptions include details regarding angles, distances, relative positions, pitch&roll and ground-contacts. Subsequently, text features are extracted from these descriptions using a language model, while a 3D human pose estimation model extracts pose features. Aligning the pose features with the text features allows for a more targeted optimization of the estimation model. Therefore, we systematically introduce three alignment approaches to effectively align features extracted by two models operating in entirely different domains. Our method incorporates prior knowledge derived from the textual descriptions into the estimation model and can be seamlessly applied to various existing framework. Experimental results on the Human3.6M and MPI-INF-3DHP datasets demonstrate that our method surpasses state-of-the-art methods.

Introduction

3D Human Pose Estimation (3DHPE) is a fundamental and important task in computer vision, which aims at predicting 3D coordinates from images or detected 2D keypoints. It plays a critical role in various applications, including person re-identification, human parsing, human action recognition, and human-computer interaction, etc (2022). With significant progress in 2D keypoint detection, prevailing methods now directly infer 3D coordinates from these detected 2D keypoints (2017; 2017; 2017; 2019). However, the inherent depth ambiguity remains a significant challenge, since a single 2D pose may correspond to multiple 3D poses. To address this, existing approaches typically incorporate two types of prior knowledge: temporal consistency and body structure representation.

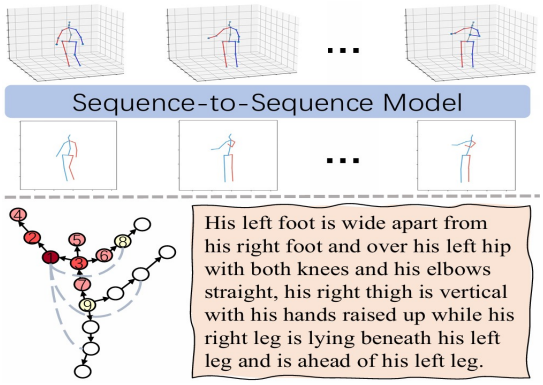

Figure 1: Three types of prior knowledge are illustrated, the upper part highlights temporal consistency, while the lower part represents body structure representation and pose textual descriptions (ours) from left to right, respectively.

Leveraging temporal consistency allows models to extract critical clues from adjacent frames, such as gait cycle, motion range of joints, etc, as illustrated in the upper part of Figure 1. A straightforward method predicts the 3D pose of an intermediate frame from a 2D pose sequence (2016), but involves redundant computations. To decrease computational overhead, many methods adopt sequence-to-sequence models to estimate the entire 3D pose sequence. Long Short-Term Memory (LSTM) (1997) models effectively retain valuable information from historical frames via memory and forgetting mechanisms (2017; 2018; 2018). However, their lack of parallelism reduces training efficiency. 3D Temporal Convolutional Networks (2018) model both spatial and temporal dependencies locally (2019; 2019), but fail to capture relationships between non-adjacent joints. Spatial and Temporal Transformers (2021) perform global modeling across entire pose sequences, effectively capturing synergistic relationships between any pair of joints, regardless of spatial or temporal distance (2021; 2023; 2023; 2022; 2024; 2024; 2022), thereby significantly enhancing representational capacity.

The other approach aims to develop a representation closely aligned with body structure, which is displayed in the lower left part of Figure 1 . Human joints and bones can

*Corresponding author.

be naturally modeled as a graph, where spatial information is effectively derived from adjacent nodes via Graph Convolutional Networks (GCNs) (2016). Some methods assign distinct weight matrices to different nodes using a weight non-sharing strategy (2019), effectively enhancing the representation capacity of GCNs. To balance the significant computational cost while maintaining the representation capacity, it is advantageous to learn a shared weight matrix along with node-specific vectors (2021). Additionally, certain methods group nodes by hop distance, allowing the model to incorporate multi-hop interactions within a single convolution layer (2021; 2023; 2021), thus mitigating information loss across network layers. More detailed methods incorporate constraints related to joint movement ranges (2018) and the inherent rigidity of bones during motion (2019) to ensure biomechanically plausible 3D pose predictions.

Although the above methods leverage human body priors, such knowledge lacks detail and requires the network to implicitly learn instance-specific features. In contrast, textual descriptions provide explicit and detailed spatial relationships for each pose. By learning the semantics, models can more effectively resolve depth ambiguity and improve 3DHPE accuracy. Therefore, we propose a 3DHPE method that incorporates knowledge from textual descriptions. Specifically, we first employ an automatic captioning pipeline (2022) to generate detailed textual descriptions for a 3D pose sequence, outlining spatial relationships among body parts, as illustrated in the lower right of Figure 1. This enriched prior knowledge serves to mitigate depth ambiguity. Subsequently, a language model extracts textual features from these descriptions, while the 3DHPE model extracts pose features. To bridge the two modalities, we introduce three feature alignment approaches that map the text features into the pose feature space. By minimizing the discrepancy between the pose features and the mapped features, we effectively inject prior knowledge into the 3DHPE model.

A few recent 3DHPE methods have mentioned textual information (2023; 2024). ActionPrompt (2023) learns text prompts based on pose classes but overlook the spatial relationships between body parts. Chatpose (2024) focuses on generating instructed poses rather than accurate joint estimation, resulting in significantly lower performance than existing 3DHPE methods. Thus, these works fall outside the scope of our study, which emphasizes leveraging fine-grained spatial semantics for accurate 3DHPE.

Generally, our contributions are summarized as follows:

- We propose a 3DHPE method that learns knowledge from textual descriptions, effectively alleviating the issue of depth ambiguity. Moreover, the abundant textual descriptions offer opportunities for further exploration. To the best of our knowledge, this is the first multi-modality 3DHPE method.
- We introduce three feature alignment techniques that bridge the pose and semantic feature spaces, enabling more effective utilization of prior knowledge. Furthermore, these techniques can be seamlessly integrated into other methods as a plug-and-play component.
- Our method outperforms existing state-of-the-art methods on the Human3.6M and MPI-INF-3DHP datasets.

Related Work

This section reviews two types of prior knowledge in 3DHPE: temporal consistency and body structure representation, and compares them with our method based on pose textual descriptions.

Temporal Consistency-Based Method

Temporal consistency constraints that 3DHPE in each frame relies on several preceding and succeeding 2D pose frames, effectively mitigating depth ambiguity. Unlike previous methods that estimated poses frame by frame and then connected them in a post-processing step, *Tekin et al* (2016) directly regressed the 3D pose in the central frames from a spatio-temporal volume, enabling end-to-end optimization. Although this method effectively enhanced estimation accuracy, it entailed significant computational costs, prompting the development of sequence-to-sequence methods. Given the LSTM network's proficiency in handling sequential problems, *Lin et al.* (2017) proposed a network composed of layer-normalized LSTM units with shortcut connections attaching the input to the output on the decoder side and imposed temporal smoothness constraint during training. *Hossain et al.* (2018) proposed a propagating LSTM network, where each LSTM unit is sequentially linked, estimating the 3D depth from the centroid to peripheral joints through learning the intrinsic joint interdependency. *Lee et al.* (2018) presented a Recurrent 3D Pose Sequence Machine, which automatically learns image-dependent structural constraint and sequence-dependent temporal context through a multi-stage sequential refinement.

When processing a frame in a sequence, LSTM networks must wait for all preceding frames to be processed, which significantly limits parallelism. In contrast, 3D temporal convolutional networks process all frames in parallel, thereby greatly enhancing training efficiency. *Pavllo et al.* (2019) employed a fully convolutional model based on dilated temporal convolutions over 2D keypoints for 3DHPE. They also introduced a semi-supervised training method known as 'back-projection' that leverages unlabeled video. *Cheng et al.* (2019) fed incomplete 2D keypoints, instead of complete but incorrect ones, to the 3D temporal convolutional network, thereby mitigating the network's susceptibility to the error-prone estimations of occluded keypoints.

3D temporal convolutional networks focus on local information within sequences, making it challenging to capture associations between non-adjacent nodes. Benefiting from the global modeling capability of Vision Transformers (ViT), ViT-based 3DHPE further enhances estimation accuracy. *Zheng et al.* (2021) designed a spatio-temporal transformer to comprehensively model the human joint relations within each frame as well as the temporal correlations across frames. To alleviate the huge computational burden caused by increased frame numbers, *Zhao et al.* (2023) leveraged a compact representation of lengthy skeletal sequences in the frequency domain to effectively scale up the receptive field and enhance robustness against noisy 2D joint detection.

Furthermore, *Tang et al.* (2023) decomposed correlation learning into space and time, presenting a spatio-temporal criss-cross attention method. Specifically, this method slices the input features into two partitions evenly along the channel dimension, followed by performing spatial and temporal attention respectively. In contrast to the parallel modeling of spatial and temporal relationships, *Zhang et al.* (2022) proposed the Mixed Spatio-Temporal Encoder (MixSTE), where spatial and temporal Transformer blocks are utilized alternately to achieve better spatio-temporal feature encoding.

Temporal consistency interprets 3DHPE from a kinematic perspective but relies on the network to infer motion patterns from surrounding frames. In contrast, our method constrains each frame with its corresponding textual description, thereby enables the network to learn specific motion patterns more directly and effectively.

Body Structure Representation-Based Method

Body structure representation can be explicitly embedded to model the synergistic relations between adjacent joints. In vanilla graph convolutional networks (GCNs), all nodes share the same weight matrix. However, this overlooks the specificity of different nodes. *Ci et al.* (2019) discarded the weight sharing scheme by freeing all of the parameters in the weight matrix, combined the advantages of GCNs and fully connected networks, thus fully unleashing the model's representational ability. However, this increases the training parameters and reduces the efficiency. Hence, *Zou et al.* (2021) devised a weight modulation that learns distinct modulation vectors for different nodes so that the feature transformations of distinct nodes are disentangled while maintaining a small model size. Furthermore, some methods attempt to model relations between non-adjacent joints, capturing motion information across a broader range. *Zeng et al.* (2021) proposed a hop-aware hierarchical channel-squeezing fusion layer to effectively extract relevant information from neighboring nodes while suppressing undesired noise. *Zhai et al.* (2023) grouped joints by k-hop neighbors and applied a hop-wise transformer-like attention mechanism to discover latent joint synergies. *Zhu et al.* (2021) proposed a graph transformer encoder-decoder with atrous convolution to effectively extract multi-scale and long-range context.

The human body structure can also be modeled as a directed graph, giving rise to another strategy for 3DHPE. In the graph, by predicting the orientation and distance of child nodes relative to their parents (2020; 2021), the entire 3D pose can be constructed. However, such methods tend to accumulate errors from central to peripheral nodes, even if the peripheral nodes are predicted with relatively high accuracy in relation to their parent nodes. On the other hand, a post-processing step is required to construct the final 3D pose.

Like temporal consistency, body structure representation interprets 3DHPE from a dynamical perspective but remains abstract. In contrast, pose textual descriptions make this concept concrete, providing specific details for each pose, thereby enabling the network to capture finer information.

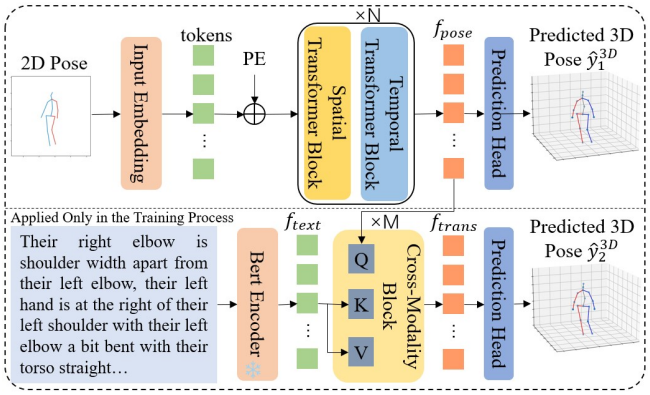

Figure 2: The framework of the proposed method.

Method

As shown in Figure 2, the proposed framework consists of three stages. First, we utilize an automatic captioning pipeline to generate comprehensive textual descriptions for a ground-truth 3D pose sequence. Subsequently, the features of the pose-text pairs are extracted separately using a pose estimation model and a text feature extraction model. To enable the pose estimation model to learn the semantic information from the textual descriptions, we introduce cross-modality blocks to align the pose features and text features. Finally, two regression heads convert the pose features and the aligned features into 3D poses, respectively. Note that predicting 3D poses using aligned features and minimizing errors serve only to constrain the features to be semantically consistent with the ground-truth poses. Therefore, the cross-modality blocks and the corresponding prediction head are used exclusively during the training process, and only the pose estimation model is required during testing.

Pose Textual Descriptions

The spatial relationships between different body parts can be described using natural language. Thanks to Ginger's work (2022), textual descriptions corresponding to ground-truth 3D poses can be directly generated through an automatic captioning pipeline. The generation process comprises several steps: given an input 3D pose, a rotation transformation is first applied to fix the coordinates, aligning the axes as shown in Figure 3. Then, posecodes are extracted based on the transformed coordinates. A posecode represents a spatial relation among a specific set of joints, encompassing five kinds of elementary relations: angles, distances, relative positions, pitch & roll, and ground-contacts. Each relation is categorized according to predefined thresholds. These posecodes are selected, aggregated, and converted to produce a textual description for the input pose.

Since the textual descriptions are only used during training, we generate them for the poses in the training sets of both the Human3.6M (2013) and MPI-INF-3DHP (2017) datasets. Figure 3 displays the poses and their corresponding textual descriptions.

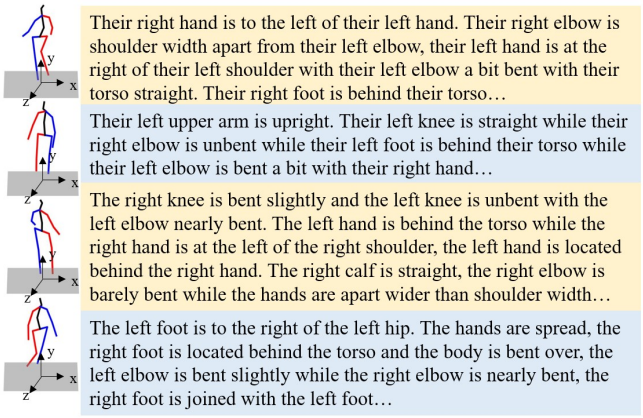

Figure 3: 3D poses and textual descriptions of Subject "S1" performing the "directions" action from the Human3.6M dataset, captured from left, front, back, and right of the subject. Particularly, the left limbs are colored blue, the right limbs red, and the torso black. The coordinate axes are oriented as follows: x-axis(left-right), y-axis(up-down), and z-axis (forward and backward).

Textual Description-Assisted 3D Human Pose Estimation

The textual descriptions of the poses delineate the spatial relationships between different body parts, thus effectively mitigating the issue of depth ambiguity. In this section, we explore how these textual descriptions assist the 3DHPE model. The model comprises two components: the pose encoder and the regression head, taking 2D poses x^{2D} as input, then extracting the pose features f_{pose}, and ultimately predicting 3D poses $\hat{y}_1^{3D}$. The estimation loss $\mathcal{L}_{err1}$ can be formulated as follows:

$$\mathcal{L}_{err1} = ||y^{3D} - \hat{y}_1^{3D}||_2 \tag{1}$$

where y^{3D} is the ground-truth 3D poses, and $||\cdot||_2$ indicates Mean Square Error (MSE) loss.

The auxiliary branch, on the other hand, takes textual descriptions as input and leverages existing pre-trained language models to extract textual features f_{text}. Note that the text encoder remains frozen during the training process, indicated by a snowflake icon. Since the pose features f_{pose} and text features f_{text} are derived from different encoders, they belong to distinct feature spaces. When calculating the distribution loss between them, it is necessary to align both within the same feature space. Due to the uncertainty in textual description generation, mapping the text features f_{text} to the pose feature space is a better choice. Additionally, by imposing a constraint that the 3D poses $\hat{y}_2^{3D}$ predicted from the aligned features f_{trans} approximate the ground-truth 3D poses, we ensure the correctness of the aligned features. The alignment loss $\mathcal{L}_{align}$ can be formulated as follows:

$$\mathcal{L}_{align} = KL(f_{pose}, f_{trans}) \tag{2}$$

where $KL(\cdot,\cdot)$ represents KL divergence loss. And the

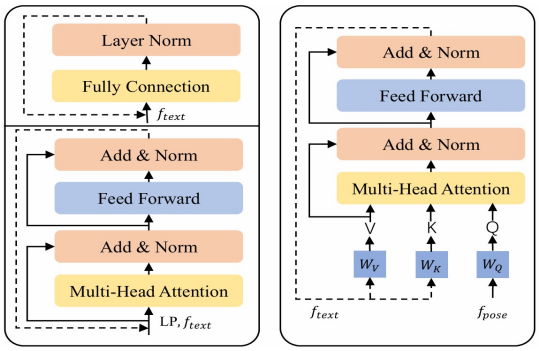

Figure 4: Three types of feature alignment approaches are arranged from left to right and top to bottom: the fully connection-based approach (FC), the transformer-based approach (TF), and the cross-modality-based approach (CM). Specifically, f_{text} represents the textual features with a shape of $f \times m \times d$, f_{pose} denotes the pose features with a shape of $f \times n \times d$, and LP indicates learnable parameters with a shape of $f \times n \times d$. Here, f, n, m, and d respectively represent sequence length, the number of keypoints, the length of the textual description, and feature dimension.

estimation loss $\mathcal{L}_{err2}$ can be defined as follows:

$$\mathcal{L}_{err2} = ||y^{3D} - \hat{y}_2^{3D}||_2 \tag{3}$$

Therefore, the total loss $\mathcal{L}_{total}$ is defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{err1} + \alpha \cdot \mathcal{L}_{align} + \beta \cdot \mathcal{L}_{err2} \tag{4}$$

where α and β are the weight coefficients.

Feature Alignment Approaches

To effectively learn the semantic information of poses from textual descriptions, we introduce three feature alignment approaches. The fully connection-based approach takes f_{text} as input and directly maps text tokens to pose tokens through a linear layer, as shown in the upper left of Figure 4. This approach fails to capture the complex relationships between the pose and text modalities, limiting the model's expressive power and restricting its ability to learn the associations between keypoints from the text. The transformer-based approach processes pose tokens into a series of learnable parameters and concatenates them with text tokens as input, which is shown in the lower left of Figure 4. This method can model complex cross-modal associations, however, the lack of guidance from pose features leads to reduced learning efficiency and performance.

Unlike fully connection-based and transformer-based approaches, this approach combines both the pose features and the text features, thereby facilitating more effective feature alignment. Specifically, this approach consists of M blocks, where the $(l+1)^{th}$ block computes the key matrix K and value matrix V based on the output w^l of the previous block, and the query matrix Q is computed based on the pose feature f_{pose} , with w^0 being f_{text}. This process is shown in the right of Figure 4 and can be formulated as follows:

MPJPE ↓	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
TCN (2019) (N=243)	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
OAN (2019) * (N=128)	38.3	41.3	46.1	40.1	41.6	51.9	41.8	40.9	51.5	58.4	42.2	44.6	41.7	33.7	30.1	42.9
PoseFormer (2021) (N=81)	41.5	44.8	39.8	42.5	46.5	51.6	42.1	42.0	53.3	60.7	45.5	43.3	46.1	31.8	32.2	44.3
P-STMO (2022) (N=81)	38.4	42.1	39.8	40.2	45.2	48.9	40.4	38.3	53.8	57.3	43.9	41.6	42.2	29.3	29.3	42.1
PoseFormerV2 (2023) (N=243)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	45.2
MixSTE (2022) (N=243)	37.6	40.9	37.3	39.7	42.3	49.9	40.1	39.8	51.7	55.0	42.1	39.8	41.0	27.9	27.9	40.9
STCFormer-L (2023) (N=243)	38.4	41.2	36.8	38.0	42.7	50.5	38.7	38.2	52.5	56.8	41.8	38.4	40.2	26.2	27.7	40.5
D3DP (2023) (N=243)	37.7	39.9	35.7	38.2	41.9	48.8	39.5	38.3	50.5	53.9	41.6	39.4	39.8	27.4	27.5	40.0
KTPFormer (2024) (N=243)	37.3	39.2	35.9	37.6	42.5	48.2	38.6	39.0	51.4	55.9	41.6	39.0	40.0	27.0	27.4	40.1
DDHPose (2024) (N=243)	37.3	40.0	35.2	37.7	41.1	46.7	38.4	38.4	52.2	53.3	41.4	38.9	38.8	27.6	27.7	39.7
LCN (2019) (N=1) *	46.8	52.3	44.7	50.4	52.9	68.9	49.6	46.4	60.2	78.9	51.2	50.0	54.8	40.4	43.3	52.7
MGCN (2021) (N=1)	45.4	49.2	45.7	49.4	50.4	58.2	47.9	46.0	57.5	63.0	49.7	46.6	52.2	38.9	40.8	49.4
HCSF w/A (2021) (N=1) *	43.1	50.4	43.9	45.3	46.1	57.0	46.3	47.6	56.3	61.5	47.7	47.4	53.5	35.4	37.3	47.9
HopFIR (2023) (N=1)	43.9	47.6	45.5	48.9	50.1	58.0	46.2	44.5	55.7	62.9	49.0	45.8	51.8	38.0	39.9	48.5
Ours (N=243, MixSTE)	37.5	39.7	36.3	38.2	41.4	48.9	38.9	38.6	50.9	55.3	41.2	39.0	39.9	27.0	27.4	40.0
Ours (N=243, D3DP)	37.6	39.3	35.3	37.8	41.4	48.7	38.9	37.7	49.7	52.5	40.7	39.2	38.9	26.7	27.0	**39.4**

Table 1: Results on the Human3.6M dataset in millimeters under MPJPE. N is the number of input frames. (*)-2D poses estimated by the Stacked Hourglass (2016), otherwise, by the Cascaded Pyramid Network (2018). The best results are highlighted in bold.

$$h^{l+1} = LN(w^l + MCA(w^l, f_{pose})) \tag{5}$$

$$w^{l+1} = LN(h^{l+1} + FF(h^{l+1})) \tag{6}$$

where $MCA(\cdot)$ represents the Multi-head Cross-Attention that calculates query, key, and value via different modalities. $LN(\cdot)$ is the Layer-Norm function to ensure the stability of the feature distribution, and $FF(\cdot)$ is the Feed-Forward module consisting of two fully connected layers and a ReLU activation function, formulated as follows:

$$FF(x) = max(0, xW_1 + b_1)W_2 + b_2 \tag{7}$$

Finally, the cross-modality feature alignment model outputs the transformed feature f_{trans}.

Experiments

To demonstrate the priority of the proposed method, we conduct experiments on two widely used datasets: the Human3.6M dataset (2013) and the MPI-INF-3DHP dataset (2017). The details are provided below.

Experimental Conditions

Human3.6M (2013) is the largest indoor dataset for 3DHPE. It contains 15 activities performed by 11 actors. Videos are captured by 4 synchronized and calibrated cameras at 50Hz. Following previous methods (2022; 2023), the 3D human pose is represented as a 17-joint skeleton, and our model is trained on 5 subjects $(S_1, S_5, S_6, S_7, S_8)$ and evaluated on 2 subjects (S_9, S_{11}). Two commonly used evaluation metrics are Mean Per Joint Position Error (MPJPE) and Procrustes MPJPE (P-MPJPE), the former computes the mean Euclidean distance between estimated and ground-truth 3D pose in millimeters, while the latter computes MPJPE after the estimated poses align to the ground truth using a rigid transformation.

MPI-INF-3DHP (2017) is a more challenging 3D pose dataset. It contains both constrained indoor scenes and complex outdoor scenes. The training set contains 8 activities performed by 8 actors from 14 camera views which cover a greater diversity of poses. The test set covers 7 activities from 6 subjects with different scenes. Following the setting in (2021; 2022; 2023), we use the valid frames provided by the official for testing. For evaluation metrics, we report MPJPE, Percentage of Correct Keypoint (PCK) within 150mm range, and Area Under Curve (AUC).

Implementation Details

We use D3DP (2023) as one of the backbones for 3DHPE, which is a diffusion model, achieving low estimation errors. The denoiser of this model is MixSTE (2022) (It is also used as the other backbone. We only report experimental results on the Human3.6M dataset because the authors only released the code on the dataset). During training, the maximum number of timesteps T is set to 1000. The model supports a customizable number of hypotheses H by sampling multiple times from a Gaussian distribution and allows for an adjustable parameter (number of iterations K) to progressively refine the final predictions during inference. Since our primary goal is to validate the efficacy of pose textual descriptions, we set both H and K to 1 in our experiments. The proposed method introduces certain enhancements to D3DP and MixSTE while retaining the default parameters and utilizing the same optimizer as in the original methods respectively. Particularly, we set α and β to 0.5 and 0.1 so that three losses are of the same order of magnitude. All experiments are conducted on GeForce RTX 3090 GPUs.

Comparison with State-of-the-art Methods

We compare methods that utilize two types of prior knowledge with the proposed method. Tables 1 and 2 display the experimental results on the Human3.6M and MPI-INF-3DHP datasets, respectively. Particularly, in Table 1, we provide detailed experimental results for various poses following existing methods.

In Table 1, we list the experimental results from top to bottom for methods based on temporal consistency, methods based on body structure representation, and our method (us-

Method	PCK ↑	AUC ↑	MPJPE ↓
TCN (2019) (N=81)	86.0	51.9	84.0
PoseFormer (2021) (N=9)	88.6	56.4	77.1
P-STMO (2022) (N=81)	97.9	75.8	32.2
MixSTE (2022) (N=27)	94.4	66.5	54.9
D3DP (2023) (N=243)	97.7	77.8	30.2
DDHPose (2024) (N=243)	98.5	78.1	29.2
DDHPose# (2024) (N=243)	97.2	77.9	30.7
Ours (N=243, D3DP)	**98.1**	**78.3**	**29.3**

Table 2: Results on the MPI-INF-3DHP dataset under three evaluation metrics using ground truth 2D keypoints. Particularly, # Indicates the results that we implement. The best results are highlighted in bold.

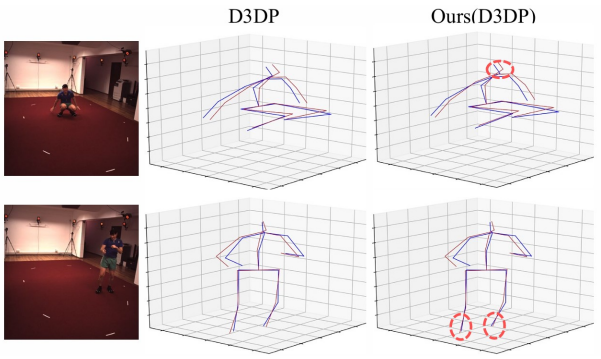

Figure 5: Visualization of predicted poses.

ing MixSTE and D3DP as baselines). Our method, "standing on the shoulders of giants", surpasses the existing methods. Additionally, we observe that methods based on temporal consistency overwhelmingly outperform those based on body structure representation. This may be because data-driven neural networks are already quite effective at learning implicit body structural information from the data. Temporal consistency further constrains the smoothness of predicted poses, and since the input includes more frames, it enhances adaptability to challenges such as depth ambiguity, occlusion, and self-occlusion.

Figure 5 displays visualizations for two poses: "sitting down" and "photo," where red denotes the predicted poses and blue indicates the ground-truth poses. Our method (using D3DP as the baseline) outperforms D3DP. In the "sitting down" pose, the predicted head position is closer to its actual position. In the "photo" pose, the predicted foot positions from our method almost coincide with the true positions. These visualizations further confirm that our method yields more accurate estimated coordinates.

Table 2 presents the experimental results on the MPI-INF-3DHP dataset. Our method (using D3DP as a baseline) demonstrates great improvement over D3DP. We reproduced the results of DDHPose and find that the outcomes are higher than those reported in the paper. Hence, we refer to our reproduced results.

Overall, our method not only demonstrates great improvements over the two baselines but also surpasses most methods based on two prior knowledge. This indicates that integrating pose textual descriptions with three feature alignment approaches effectively incorporates prior knowledge into the model.

MPJPE ↓	R1	R2	R3	R4	R5
Head	<u>35.04</u>	50.41	52.78	38.89	36.97
Left Hand	34.90	<u>50.46</u>	52.72	38.92	37.04
Right Hand	35.00	50.45	<u>52.92</u>	38.87	36.86
Left Foot	34.98	50.17	52.41	<u>39.38</u>	36.69
Right Foot	35.03	50.37	52.67	38.99	<u>37.20</u>

Table 3: Results for the corresponding regions on the Human3.6M dataset after removing descriptions related to edge nodes. The largest estimation error is underscored.

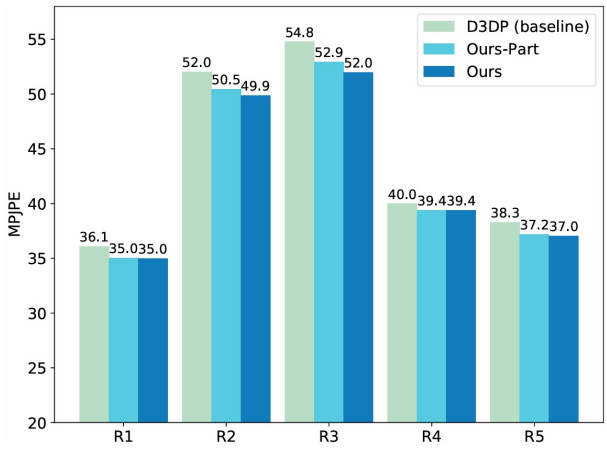

Figure 6: The errors for the corresponding regions on the Human3.6M dataset after removing descriptions related to edge nodes are compared between the baseline and our method. "Ours-Part" refers to our method with the corresponding edge node textual descriptions removed.

Ablation Studies and Analyses

We conduct ablation experiments to validate the effectiveness of the pose textual descriptions and the proposed three alignment approach.

Impact of Pose Textual Descriptions We remove the descriptions related to the head (node 10), left hand (node 13), right hand (node 16), left foot (node 6), and right foot (node 3) from the complete pose textual descriptions and train the corresponding models. Table 3 shows the results of these models in the regions related to the aforementioned edge nodes: R1 (nodes 8, 9, 10), R2 (nodes 11, 12, 13), R3 (nodes 14, 15, 16), R4 (nodes 4, 5, 6), and R5 (nodes 1, 2, 3) in the Human3.6M dataset. We observe that for any region Ri, removing the descriptions related to the edge nodes increases the estimation error for that region. This indicates that the network can enhance 3D keypoint estimation accuracy to some extent by learning the semantics in the descriptions. Moreover, we find that the estimation errors are primarily concentrated in the regions corresponding to the left and right hands, followed by those of the left and right feet, with the head region exhibiting the smallest error. The main rea-

Alignment	MPJPE↓
-	40.0
FC	39.7
TS	39.6
CM	**39.4**

(a) D3DP backbone

Alignment	MPJPE↓
-	40.9
FC	40.6
TS	40.4
CM	40.0

(b) MixSTE backbone

Table 4: Ablation study of different feature alignment approaches on the Human3.6M dataset. The best results are highlighted in bold.

Backbone	Alignment	PCK ↑	AUC ↑	MPJPE ↓
D3DP	-	97.7	77.8	30.2
D3DP	FC	97.9	78.0	29.7
D3DP	TS	97.8	77.9	29.8
D3DP	CM	**98.1**	**78.3**	**29.3**

Table 5: Ablation study of different feature alignment approaches on the MPI-INF-3DHP dataset. The best results are highlighted in bold.

son for this is that the hands have higher degrees of freedom relative to the hip joint (node 0).

Furthermore, we compare the estimation errors of the baseline method, the proposed method (Ours), and the method that removes the region edge node-related pose descriptions from the proposed method (Ours-Part) across five regions on the Human3.6M dataset, as shown in Figure 6. The experimental results indicate that the proposed method outperforms the baseline method in all regions, with the largest improvements observed in R2 and R3. These two regions, being the farthest from the hip joint, suffer the most from depth ambiguity, suggesting that textual descriptions indeed help mitigate the impact of depth ambiguity. Additionally, compared to the Ours-Part method, our method still shows the greatest improvement in R2 and R3, with limited improvements in the other regions. This indicates that descriptions related to the hands play the most significant role in enhancing keypoint recognition accuracy.

Effect of Feature Alignment Approaches Tables 4 and 5 present the experimental results on the Human3.6M and MPI-INF-3DHP datasets, respectively. We utilize two baselines, D3DP and MixSTE, and explore three different feature alignment approaches (the fully connection-based approach, the transformer-based approach, and the cross-modality-based approach) to assess their capability in leveraging pose textual descriptions. As the developers of MixSTE only released training code for the Human3.6M dataset, we provide ablation results solely for this dataset.

The results in Tables 4 and 5 demonstrate that even with several simple fully connected layers, our method effectively enhances the model's performance, indicating that pose textual descriptions indeed alleviate depth ambiguity to some extent. When employing the transformer-based approach with stronger representational power, performance improves further, although the degree of improvement diminishes. This confirms the substantial differences between the feature spaces of the pose feature extraction model and the text feature extraction model, preventing the two types of features from effectively converging. Furthermore, the cross-modality alignment approach specifically maps text features into the pose space, resulting in another boost in model performance.

Absolute difference ↓	x	y	z
D3DP	14.06	14.90	28.70
Ours (D3DP)	13.89	15.00	27.92
MixSTE	13.98	14.87	29.79
Ours (MixSTE)	13.73	14.83	28.98

Table 6: Comparison between the proposed method and baselines across three axes on the Human3.6M dataset.

In addition, when using D3DP as the baseline, our method achieves an improvement of 0.6, whereas with MixSTE as the baseline, the improvement is 0.9. The proposed method demonstrates a greater enhancement relative to both baselines but shows stronger improvement capabilities with the weaker model. This substantiates that many previous methods indeed face depth ambiguity, and that pose textual descriptions can effectively mitigate this issue.

Moreover, the experimental results in Table 5 show that the fully connection-based approach is more effective than the transformer-based approach, suggesting that the latter does not always outperform the former. The underlying reason may be that neither method fully achieves effective feature alignment. Conversely, the cross-modality-based approach shows more improvements across all three metrics, thereby demonstrating its superior capability in feature alignment.

Improvement in X, Y, and Z Axes We also compare the proposed method with the baselines in terms of absolute differences across the three axes, defined as $|\hat{c} - c|$, where $\hat{c}$ and c denote the predicted and ground-truth values along a given axis, respectively. Specifically, the z-axis represents depth. Overall, the absolute difference is largest along the z-axis, followed by the y-axis, with the smallest difference along the x-axis, as shown in Table 6. This indicates that existing methods indeed suffer from depth ambiguity. The primary improvement is observed along the z-axis, where errors are reduced by 0.78 and 0.81, respectively, while the results along the x and y axes remain comparable to those of the baselines. This demonstrates that the proposed method effectively mitigates the depth ambiguity.

Conclusion

This paper proposes a 3DHPE method capable of learning from textual descriptions. Firstly, an automatic captioning pipeline generates textual descriptions for a 3D pose sequence, which outlines the spatial relationships between different body parts. To better leverage these descriptions, we also introduce three feature alignment approach. Consequently, the model incorporates prior knowledge from the pose textual descriptions, thereby surpassing the existing methods on the Human3.6M and MPI-INF-3DHP datasets.

Acknowledgments

This work was supported by National Natural Science Foundation of China 62376255 and project from China Merchants Bank under Grant FTIT2022058.

References

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6836–6846.

Bai, S.; Kolter, J. Z.; and Koltun, V. 2018. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv:1803.01271*.

Cai, Q.; Hu, X.; Hou, S.; Yao, L.; and Huang, Y. 2024. Disentangled Diffusion-Based 3D Human Pose Estimation with Hierarchical Spatial and Temporal Denoiser. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 882–890.

Chen, C.-H.; and Ramanan, D. 2017. 3d human pose estimation= 2d pose estimation+ matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7035–7043.

Chen, C.-H.; Tyagi, A.; Agrawal, A.; Drover, D.; Mv, R.; Stojanov, S.; and Rehg, J. M. 2019. Unsupervised 3d pose estimation with geometric self-supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5714–5724.

Chen, T.; Fang, C.; Shen, X.; Zhu, Y.; Chen, Z.; and Luo, J. 2021. Anatomy-aware 3d human pose estimation with bone-based pose decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1): 198–209.

Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; and Sun, J. 2018. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7103–7112.

Cheng, Y.; Yang, B.; Wang, B.; Yan, W.; and Tan, R. T. 2019. Occlusion-aware networks for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF international conference on computer vision*, 723–732.

Ci, H.; Wang, C.; Ma, X.; and Wang, Y. 2019. Optimizing network structure for 3d human pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2262–2271.

Delmas, G.; Weinzaepfel, P.; Lucas, T.; Moreno-Noguer, F.; and Rogez, G. 2022. Posescript: 3d human poses from natural language. In *European Conference on Computer Vision*, 346–362. Springer.

Feng, Y.; Lin, J.; Dwivedi, S. K.; Sun, Y.; Patel, P.; and Black, M. J. 2024. Chatpose: Chatting about 3d human pose. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2093–2103.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Hossain, M. R. I.; and Little, J. J. 2018. Exploiting temporal information for 3d human pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, 68–84.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2013. 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.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7122–7131.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Lee, K.; Lee, I.; and Lee, S. 2018. Propagating lstm: 3d pose estimation based on joint interdependency. In *Proceedings of the European conference on computer vision (ECCV)*, 119–135.

Lin, M.; Lin, L.; Liang, X.; Wang, K.; and Cheng, H. 2017. Recurrent 3d pose sequence machines. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 810–819.

Liu, W.; Bao, Q.; Sun, Y.; and Mei, T. 2022. Recent advances of monocular 2d and 3d human pose estimation: A deep learning perspective. *ACM Computing Surveys*, 55(4): 1–41.

Martinez, J.; Hossain, R.; Romero, J.; and Little, J. J. 2017. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, 2640–2649.

Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; and Theobalt, C. 2017. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, 506–516. IEEE.

Newell, A.; Yang, K.; and Deng, J. 2016. 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*, 483–499. Springer.

Pavllo, D.; Feichtenhofer, C.; Grangier, D.; and Auli, M. 2019. 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*, 7753–7762.

Peng, J.; Zhou, Y.; and Mok, P. 2024. KTPFormer: Kinematics and Trajectory Prior Knowledge-Enhanced Transformer for 3D Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1123–1132.

Shan, W.; Liu, Z.; Zhang, X.; Wang, S.; Ma, S.; and Gao, W. 2022. P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In *European Conference on Computer Vision*, 461–478. Springer.

Shan, W.; Liu, Z.; Zhang, X.; Wang, Z.; Han, K.; Wang, S.; Ma, S.; and Gao, W. 2023. Diffusion-based 3d human pose

estimation with multi-hypothesis aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14761–14771.

Tang, Z.; Qiu, Z.; Hao, Y.; Hong, R.; and Yao, T. 2023. 3D Human Pose Estimation With Spatio-Temporal Criss-Cross Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4790–4799.

Tekin, B.; Rozantsev, A.; Lepetit, V.; and Fua, P. 2016. Direct prediction of 3d body poses from motion compensated sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 991–1000.

Tome, D.; Russell, C.; and Agapito, L. 2017. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2500–2509.

Xu, J.; Yu, Z.; Ni, B.; Yang, J.; Yang, X.; and Zhang, W. 2020. Deep kinematics analysis for monocular 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on computer vision and Pattern recognition*, 899–908.

Zeng, A.; Sun, X.; Yang, L.; Zhao, N.; Liu, M.; and Xu, Q. 2021. Learning skeletal graph neural networks for hard 3d pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11436–11445.

Zhai, K.; Nie, Q.; Ouyang, B.; Li, X.; and Yang, S. 2023. HopFIR: Hop-wise GraphFormer with Intragroup Joint Refinement for 3D Human Pose Estimation. *arXiv preprint arXiv:2302.14581*.

Zhang, J.; Tu, Z.; Yang, J.; Chen, Y.; and Yuan, J. 2022. 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*, 13232–13242.

Zhao, Q.; Zheng, C.; Liu, M.; Wang, P.; and Chen, C. 2023. PoseFormerV2: Exploring Frequency Domain for Efficient and Robust 3D Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8877–8886.

Zheng, C.; Zhu, S.; Mendieta, M.; Yang, T.; Chen, C.; and Ding, Z. 2021. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11656–11665.

Zheng, H.; Li, H.; Shi, B.; Dai, W.; Wang, B.; Sun, Y.; Guo, M.; and Xiong, H. 2023. Actionprompt: Action-guided 3d human pose estimation with text and pose prompting. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 2657–2662. IEEE.

Zhu, Y.; Xu, X.; Shen, F.; Ji, Y.; Gao, L.; and Shen, H. T. 2021. PoseGTAC: Graph Transformer Encoder-Decoder with Atrous Convolution for 3D Human Pose Estimation. In *IJCAI*, 1359–1365.

Zou, Z.; and Tang, W. 2021. Modulated graph convolutional network for 3D human pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11477–11487.