SVTformer: Spatial-View-Temporal Transformer for Multi-View 3D Human Pose Estimation

Wanruo Zhang[1], **Mengyuan Liu**[1*], **Hong Liu**[1], **Wenhao Li**[2]

[1]State Key Laboratory of General Artificial Intelligence, Peking University, Shenzhen Graduate School
[2]Nanyang Technological University
wanruo_zhang@stu.pku.edu.cn; {liumengyuan, hongliu}@pku.edu.cn; wenhao.li@ntu.edu.sg

Abstract

Recently, transformer-based methods have been introduced to estimate 3D human pose from multiple views by aggregating the spatial-temporal information of human joints to achieve the lifting of 2D to 3D. However, previous approaches cannot model the inter-frame correspondence of each view's joint individually, nor can they directly consider all view interactions at each time, leading to insufficient learning of multi-view associations. To address this issue, we propose a Spatial-View-Temporal transformer (SVTformer) to decouple spatial-view-temporal information in sequential order for correlation learning and model dependencies between them in a local-to-global manner. SVTformer includes an attended Spatial-View-Temporal (SVT) patch embedding to attentively capture the local features of the input poses and stacked SVT encoders to extract global spatial-view-temporal dependencies progressively. Specifically, SVT encoders perform three reconstructions sequentially to attended features with the learning through view decoupling for temporal-enhanced spatial correlation, temporal decoupling for spatial-enhanced view correlation, and another view decoupling for spatial-enhanced temporal relationship. This decoupling-coupling-decoupling multi-view scheme enables us to alternatively model the inter-joint spatial relationships, cross-view dependencies, and temporal motion associations. We evaluate the proposed SVTformer on three popular 3D HPE datasets, and it yields state-of-the-art performance. It effectively deals with ill-posed problems and enhances the accuracy of 3D human pose estimation.

Code — https://github.com/Rowenazhang/SVTformer

Introduction

3D human pose estimation (HPE) is essential in computer vision. It aims to estimate the 3D coordinates of human joints in images or videos to reconstruct human posture. It is widely used in many fields, such as human action recognition (Rajasegaran et al. 2023), human motion prediction (Wang et al. 2023b), person re-identification (Wang et al. 2022) and so on.

Monocular 3D HPE (Li et al. 2022; Yu et al. 2023; Li et al. 2023a; Wang et al. 2024) refers to estimating the positions of 3D human joints from a single image or monocular video. However, the issues of depth ambiguity and partial occlusion in the single-view setting hinder more accurate 3D HPE. A natural way to solve them is to use multi-view information for 3D HPE (Jiang, Hu, and Xia 2023; Zhang et al. 2022b; Wandt et al. 2021), which estimates human poses in 3D space by using data from multiple camera views.

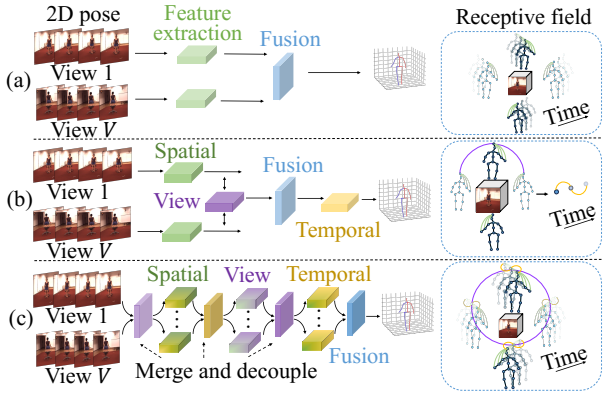

Figure 1: Comparison between previous multi-view 3D HPE methods and ours. (a) is the CNN/RNN/GCN-based method with only the spatial receptive field. (b) is the transformer-based method with spatial and adjacent view receptive fields and the temporal receptive field after view fusion. (c) is our SVTformer, incorporating all spatial, view, and temporal receptive fields. Since our method models richer spatial, view, and temporal correlations sequentially in a local-to-global manner through continuous reshaping and attended learning of the input 2D pose, our method outperforms others.

The straightforward way to achieve multi-view 3D HPE (Ma et al. 2021; Shuai, Wu, and Liu 2022; Hua et al. 2022) is generally to first estimate the 2D pose of the human body in each view, then use the 2D skeleton of different views as the input of the feature extraction network for 2D-3D lifting, and fuse the multi-view information at a deeper stage to obtain the 3D pose, as shown in Fig. 1(a). Its receptive field is mainly confined to a single view. An off-the-shelf 2D pose estimator (Chen et al. 2018) is often used to detect the 2D pose of the human body. Our work mainly focuses on the progress of recovering 3D human poses from the obtained multi-view 2D poses. This involves addressing the correlation between joint positions corresponding to dif-

*Corresponding author

ferent camera perspectives to fuse information from multiple viewpoints effectively.

Many multi-view fusion methods divide multiple views into multiple different monocular 2D to 3D HPE procedures and fuse them at the final stage. Several efforts have been made to solve the correlation issues of multiple views through the temporal association of different views (Chu et al. 2021), epipolar geometry (Qiu et al. 2019; Wang et al. 2023a), and multi-view consistency constraints (Rhodin et al. 2018; Kim et al. 2022; Wan, Chen, and Zhao 2023). However, these models almost only consider the temporal relationship of joints under a single view or the association across views, lacking the coherent learning of the spatial-temporal information between joints under each view respectively and the mining of spatial information across all views at the same time. Secondly, most of these methods only fuse cross-view features in the last step but ignore the multi-view information in the shallow part. Furthermore, previous work often relies on specific camera parameters and requires complex constraints for cross-view learning, making it difficult to implement the model in real scenarios.

Existing transformer-based methods (Shuai, Wu, and Liu 2022) directly merge the temporal input of each view into a vector but ignore the different spatial relationships and temporal correspondence of the joints in each view. In single-view 3D HPE, researchers (Zhao et al. 2019) propose that GCN can extract and process the joint spatial relationship. Later, in multi-view 3D HPE, some transformer-based methods (Zhou et al. 2023; Zhang et al. 2024) introduce graph structures into transformers to enhance the local expression of spatial features in each view. As shown in Fig. 1(b), firstly, the spatial relationship between the joints in each view is fused with attention. Then, the features of the two adjacent views are interacted and enhanced through the cross-view attention model. Finally, the multi-view features are embedded through attention-based temporal feature extraction. This type of method only considers the feature fusion between two adjacent views but does not explicitly mine the correspondence among all views. In addition, this strategy only considers the temporal relationship of the unified sequence after the fusion process of multi-view spatial information without exploring the temporal motion of the human joints in each original view before fusion. Furthermore, graph-based methods represent features locally and insufficiently, which limits the receptive field of the model and introduces additional structural priors of the joint connection, affecting the generalization ability of the model. Therefore, this inspires us to further explore the associations and fusion methods of multiple views, time, and space.

In order to improve the extraction of short-term and long-term correlation and the fusion of spatial-temporal-view information in multiple views, we propose a Spatial-View-Temporal transformer (SVTformer), an effective multi-view fusion method for 3D HPE with sequential spatial-view-temporal attention in a local-to-global manner. It merges the rich spatial, multi-view, and temporal correlations by constantly decoupling different views and frames to reshape the input pose. SVTformer not only extracts local spatial-view-temporal correlations in the shallow layer but also alternately captures their global dependencies in deeper networks.

As shown in Fig. 1(c), our SVTformer considers correlation learning by sequentially decoupling it into three aspects: space, multi-view, and time. Specifically, to capture the spatial correlation of joints in the same view, it first merges the input features and decouples multiple views. It then constructs a spatial attention model enhanced with temporal information to integrate the spatial correlation of different joints in each view. Next, considering the spatial correlation among different views at each time step, we merge and decouple the inputs again to separate each frame and utilize the view attention enhanced by spatial information to obtain cross-view features. Finally, considering the temporal relationship of joints under each view, we decouple different viewpoints again after merging and fuse the temporal correlation within each view through spatially enhanced temporal attention. By stacking the alternate learning of these three attentions, a more complete and rich multi-view spatial-temporal fusion is achieved. In particular, SVTformer includes an input embedding module, named Attended SVT Patch Embedding, based on spatial-view-temporal attention in the shallow layer to extract local features related to space, view, and time in the original input for subsequent deeper multi-view spatial-temporal correlation learning.

Our main contributions are as follows:

- We propose an effective SVTformer framework based purely on transformers, which sequentially decouples the multi-view features in 3D HPE into spatial, view, and temporal correlations.
- By local-to-global fusion, SVTformer extracts short-range relationships of spatial-view-temporal features through attended patch embedding and gradually explores the long-range correlations in an alternating and sequential manner, effectively alleviating the depth ambiguity problem and improving the estimation accuracy of 3D HPE without the requirement for extra camera parameters, complex constraints, and human priors.
- Extensive experimental results on three popular datasets illustrate that our SVTformer performs better than existing 3D HPE approaches in indoor and outdoor scenes.

Related Work

Multi-view 3D HPE. Multi-view 3D HPE methods have the potential to enhance depth perception and address issues caused by occlusion and depth ambiguity. Existing multi-view methods (Kadkhodamohammadi and Padoy 2021; He et al. 2020; Qiu et al. 2019) infer 3D poses based on triangulation of camera calibration and detected 2D poses, relying heavily on accurate camera parameters with restricted generalization performance. Therefore, subsequent work (Gordon et al. 2022) only learns 3D rotations between skeleton parts and bone lengths that are independent of camera position. Besides, other approaches (Wan, Chen, and Zhao 2023; Kim et al. 2022, 2024) take multi-view consistency as supervision in 3D HPE to alleviate the problem of independence between multi-view 2D poses and refine the estimation error. Differently, we construct a transformer-based model that

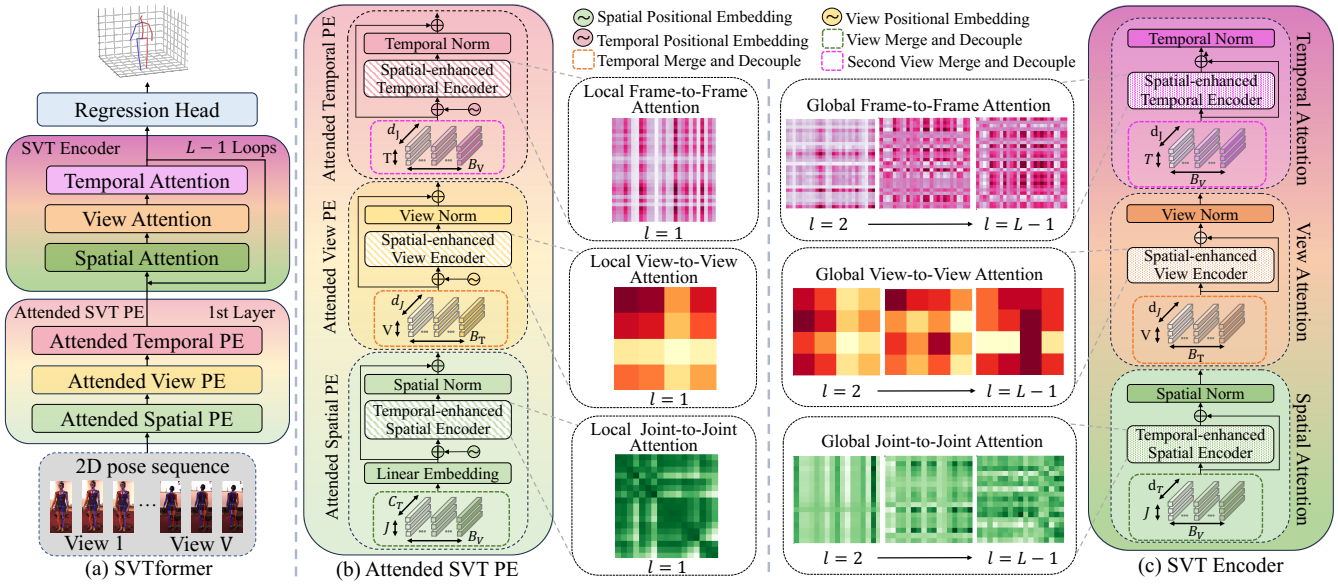

Figure 2: (a) The Overview of our SVTformer in a sequential alternating manner with (b) an attended spatial-view-temporal patch embedding layer and (c) stacked spatial-view-temporal encoders. The vanilla transformer encoder to different inputs is applied as our encoder. The attention maps for spatial, view, and temporal associations are visualized for each layer, with darker colors indicating stronger attention. The attention maps in the shallow layer with attended SVT PE exhibit diagonal structures or short-range relationships, indicating a stronger emphasis on local correlation modeling. The attention maps in the deeper layers with the SVT encoder become sparser as the layers deepen, suggesting a greater focus on global relationship modeling.

pays attention to time, space, and view separately. By alternately learning the attention in these three parts in order, we can capture the correlation of inter-joints, multiple views, and human temporal motions without relying on any camera parameters and complex constraints.

Transformer-Based 3D HPE. The transformer has a strong ability to model long sequences and capture global information. In single-view 3D HPE, many approaches (Tang et al. 2023; Zhang et al. 2022a; Li et al. 2024; Zhai et al. 2023) use transformers to learn spatial-temporal correlations. Multi-view 3D HPE mainly concentrates on aggregating information from 2D joints, frame series, and multiple views to regress 3D poses. Only a few multi-view methods deploy transformers to mine the spatial-temporal relationships. MTF-Transformer (Shuai, Wu, and Liu 2022) applies transformers to fuse multi-view and temporal information, respectively, but introduces additional 2D pose confidence and more complex loss functions as model constraints. HMV-former (Zhou et al. 2023) and SGraFormer (Zhang et al. 2024) incorporate the prior spatial information of human joints into the transformer with graph representation. However, they only consider the relationships between adjacent views and the temporal dependencies of fused features. Our method correlates the temporal and view information of human joints in all input views before the final feature fusion, and the correlation learning is carried out in a sequential order of space-view-time, which deeply mines more comprehensive correspondence in multiple views.

Method

For multi-view 3D HPE, our input is an image sequence $\mathfrak{I} = \{I_i\}_{i=1}^{B\times\mathcal{V}\times T}$ containing T frames and $\mathcal{V}$ views, and B samples, as shown in Fig. 2. We use an off-the-shelf 2D pose estimator (Chen et al. 2018) to obtain the 2D joint coordinates $\mathbb{C}_{T,\mathcal{V},J} \in \mathbb{R}^{B\times T\times\mathcal{V}\times J\times 2}$ of human body from T frames across $\mathcal{V}$ views with J joints and 2 channels. The obtained 2D pose is further input into the 2D-3D lifting network, and regress the target 3D pose $P_{T,J} \in \mathbb{R}^{T\times J\times 3}$ corresponding to the multi-view image sequence through the regression head. In our SVTformer, we first propose attended spatial-view-temporal (SVT) patch embedding to attentively capture the local detailed information of the input poses. Then, to investigate the global dependencies in SVT features, a stacked SVT encoder is built for fusion sequentially.

The Fundamentals of Transformer Encoder

The transformer encoder mainly consists of a Multi-head Self-Attention module (MSA) and a Feed-Forward Network (FFN). By first linearly mapping the token $X \in \mathbb{R}^{N\times C}$ containing C channels and N inputs to form three matrices $Q, K, V \in \mathbb{R}^{N\times d}$, representing queries, keys, and values respectively, one transformer encoder can be expressed as

$$\begin{aligned} MSA(Q,K,V) &= Softmax(\frac{QK^T}{\sqrt{d}})V, \\ FFN(X') &= GeLU(X'\cdot W_1)\cdot W_2, \\ X' &= MSA(LN(X)) + X, \\ Y &= FFN(LN(X')) + X', \end{aligned} \tag{1}$$

where d is the key's dimension, W_1 and W_2 are two projection matrices, and $LN(\cdot)$ means layer normalization.

Attended SVT Patch Embedding

Traditional transformers use patch embedding to convert image data into a series of fixed-length vectors for subsequent transformer processing. This part often involves dividing the input image into small patches of uniform size, flattening each patch, and mapping it to a high-dimensional space through a linear projection layer. Different from traditional methods, in our model, in order to better convert the input 2D skeleton sequence into an embedded representation suitable for our SVTformer processing, our Attend SVT Patch Embedding (Attended SVT PE) uses a transformer encoder structure to sequentially pass the input 2D posture sequence through a layer of joint point space, multi-view, and temporal correlation attention, thereby mapping it to a high-dimensional space with attention.

Attended Spatial Patch Embedding. First, in order to enhance the spatial association between the joints of each view, we reshape the input, including merging the sample size and channel number of the 2D pose sequence $\mathbb{C}_{T,\mathcal{V},J}$, and decoupling different views. Specifically, V of the B samples are merged into B_V as the new view sample size, where $B_V = B \cdot V$. Meantime, the number of frames and channels is merged to obtain the new number of channels enhanced by temporal features, expressed as $C_T = C \cdot T$ with the input 2D coordinate dimension $C = 2$. In this way, for the reshaped temporal enhanced features $\mathbb{C}_s \in \mathbb{R}^{B_V \times J \times C_T}$, we project the features of each joint point into a high-dimensional feature $X_s \in \mathbb{R}^{B_V \times J \times d_T}$ through a linear layer, where $d_T = d \times T$. Then, we use the positional embedding matrix to preserve the spatial location information, represented as PE_s. Our temporal-enhanced spatial transformer encoder (TS-E) uses a vanilla transformer encoder layer to separate different views in the spatial joint dimension so that all joints of each view form a spatial token $o \in \mathbb{R}^{1 \times J \times d_T}$, and the spatial information of joints of different views is modeled in parallel. Specifically, the temporal-enhanced spatial token of each view is fed into the transformer encoder to compute spatial attention, and the enhanced spatial patch embedding is obtained by performing layer normalization (LN) along spatial dimension and residual connection. Our attended spatial patch embedding $Y_s \in \mathbb{R}^{B_V \times J \times d_T}$ is formulated as follows:

$$Y_s = LN(TS\text{-}E(X_s + PE_s)) + X_s. \tag{2}$$

Attended View Patch Embedding. Then, in order to enhance the correspondence of joints across views at each moment, we constructed attended view patch embedding. Similarly, the output Y_s of attended spatial patch embedding is reshaped, which merges into a new batch B_T and the number of channels d_J by decoupling at different moments, and then spatially-enhanced view features X_v are obtained, denoted as $Y_s \in \mathbb{R}^{B_V \times J \times d_T} \rightarrow X_v \in \mathbb{R}^{B_T \times V \times d_J}$, $\rightarrow$ denoting reshape operation, $B_T = B \times T, d_J = d \times J$. We embed the view positional matrix into the input to retain the independent positional information of each view. Through our spatial-enhanced view transformer encoder (SV-E), we separate different moments in the view dimension and use the transformer encoder to model the correspondence between the joints of multiple views at each moment so that we can capture the cross-view correlation. Similarly, we perform layer normalization along the view dimension and residual connection on the view feature dimension to obtain our attended view patch embedding, expressed as $Y_v \in \mathbb{R}^{B_T \times V \times d_J}$:

$$Y_v = LN(SV\text{-}E(X_v + PE_v)) + X_v. \tag{3}$$

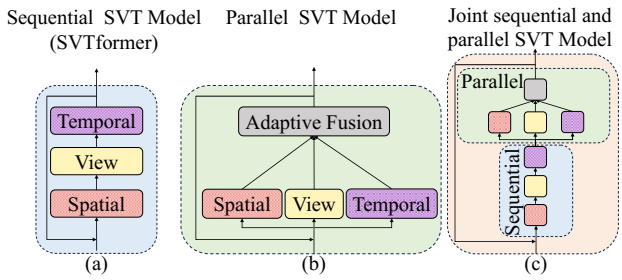

Figure 3: Three variants of our SVT transformer.

Attended Temporal Patch Embedding. After obtaining the cross-view enhanced features, we further learn the effective temporal motion of the joints. In order to enhance the association between joints in different frames under each view, we pass attended view patch embedding through batch-view and spatial-channel merges and view decoupling to obtain spatially-enhanced temporal features X_t, denoted as $Y_v \in \mathbb{R}^{B_T \times V \times d_J} \rightarrow X_t \in \mathbb{R}^{B_V \times T \times d_J}$. Then, combined with the temporal positional embedding to consider the positional information of each frame, we use the residual and layer normalization on the temporal dimension to obtain our attended temporal patch embedding $Y_t \in \mathbb{R}^{B_V \times T \times d_J}$. The main calculations of our spatially-enhanced temporal encoder (ST-E) are as follows:

$$Y_t = LN(ST\text{-}E(X_t + PE_t)) + X_t. \tag{4}$$

Spatial-View-Temporal Transformer

Attended SVT PE encodes the relationships between short-range information in the input poses. We further explore the global relationships between long-range spatial-view-temporal correlations. Based on the three encoders discussed above, i.e., TS-E, SV-E, and ST-E, we develop three variants of the spatial-view-temporal transformer.

Sequentially Alternating Model. After attending to the SVT patch embedding, we use a stacked SVT encoder to further learn the spatial dependency, cross-view correspondence and temporal motion of different input joints. To effectively model the multi-view 2D pose sequence in series, we design a sequentially alternating spatial-view-temporal transformer structure, which our SVTformer follows, as shown in Fig. 2. Similar to attended SVT patch embedding, we correspondingly propose three attention modules, as shown in Fig. 3(a), namely spatial attention to model the spatial relationship using temporal enhancement in each

view, view attention to learn the association across views, and temporal attention to model different temporal motions in the same view. For each attention module, the input is reshaped through merge and decouple operations and then fed into the encoder module for enhanced feature representation. The difference from attended SVT patch embedding is that our spatial, view, and temporal attentions in SVT encoders directly connect the input and output of the transformer encoder with residual connections and then perform layer normalization. This method can better normalize the output of encoders and residuals and reduce the changes in the input distribution of each attention module, which helps to stabilize the gradient propagation and improve the convergence speed of the model. Secondly, they no longer contain the positional embedding step. We express the alternating learning process through the lth layer as:

$$\begin{aligned} Y_t^{l-1} \in \mathbb{R}^{B_V \times T \times d_J} &\rightarrow X_s^l \in \mathbb{R}^{B_V \times J \times d_T}, \\ Y_s^l &= LN(TS\text{-}E(X_s^l) + X_s^l), \\ Y_s^l \in \mathbb{R}^{B_V \times J \times d_T} &\rightarrow X_v^l \in \mathbb{R}^{B_T \times V \times d_J}, \\ Y_v^l &= LN(SV\text{-}E(X_v^l) + X_v^l), \\ Y_v^l \in \mathbb{R}^{B_T \times V \times d_J} &\rightarrow X_t^l \in \mathbb{R}^{B_V \times T \times d_J}, \\ Y_t^l &= LN(ST\text{-}E(X_t^l) + X_t^l), \end{aligned} \tag{5}$$

where $l = 1, 2, ..., L$, $Y_t^0 = Y_t$.

Parallel Weighted Model. As shown in Fig. 3(b), we first transform the 2D skeleton joints $\mathbb{C}_{T,\mathcal{V},J} \in \mathbb{R}^{B \times T \times \mathcal{V} \times J \times 2}$ through our reshape operation including merge and decouple to form the input of spatial attention $X_s \in \mathbb{R}^{B_V \times J \times C_T}, C_T = C \times T$, the input of view attention $X_v \in \mathbb{R}^{B_T \times V \times C_J}, C_J = C \times J$ and the input of temporal attention $X_t \in \mathbb{R}^{B_V \times T \times C_J}$. The three branches map the input to high dimensions through linear embedding and embed it with positional embedding, respectively, to obtain the input of spatial, view, and temporal transformer encoder. Through the calculation of residual and attention, the features containing spatial, view, and temporal attention are obtained, which are $\{Y_s^l, Y_v^l, Y_t^l\} \in \mathbb{R}^{B \times d \times T \times V \times J}$ respectively. We design an adaptive fusion method to weigh and fuse the three branches. Specifically, for each attention output, we reduce its feature map to $B \times d \times 1 \times 1 \times 1$ through global max pooling along the temporal, view, and spatial dimensions. The three sets of features are concatenated and fed into two fully connected layers to capture the contextual information among d channels and three branches, respectively. After a softmax normalization, the adaptive weights of three branches are formed as $\alpha_s, \alpha_v, \alpha_t$. For each transformer layer, we aggregate the three attention outputs with the weighted summation:

$$Y_{svt}^l = Y_s^l \cdot \alpha_s + Y_v^l \cdot \alpha_v + Y_t^l \cdot \alpha_t, \quad l = 1, 2, ..., L. \tag{6}$$

Joint Sequential and Parallel Model. We also explore combining sequential and parallel structures to learn the associations in three dimensions: time, space, and view, shown in Fig. 3(c). We first perform sequential embedding to learn the spatial-view-temporal associations of 2D joints and then feed them into the parallel SVT transformer model to adaptively learn the effective representations of the three at the same time. We can also pass the features through the parallel SVT transformer first, after which they are fed into the sequential transformer, expressed with *.

Experiments

Datasets and Protocols

Human3.6M. (Ionescu et al. 2013) is the most widely-used 3D HPE dataset. It contains 3.6 million 3D human poses and images captured from 4 cameras. This dataset consists of 11 actors performing 15 daily activities in an indoor laboratory. Subjects 1,5,6,7,8 are often used for training, and subjects 9,11 are for testing. Two standard evaluation protocols are used to verify the effectiveness: Protocol 1 (P1) calculates the Mean Per Joint Position Error (MPJPE) between the estimated pose and ground truth, and Protocol 2 (P2) calculates the Procrustes-MPJPE using the MPJPE after rigid alignment.

MPI-INF-3DHP. (Mehta et al. 2017) is a large-scale 3D HPE dataset containing 8 actors performing in three scenes, i.e., green-screen, non-green screen, and outdoor, with 1.3 million frames captured by 14 cameras. We use the four main views of subjects S1-S6 as training sets and S7 and S8 for testing. The evaluation indicators of this dataset are P1, P2, Percentage of Correct Keypoints (PCK) with a threshold of 150 mm, and corresponding Area Under Curve (AUC).

Ski-Pose PTZ-Camera. (Fasel et al. 2016) is a more challenging 3D HPE dataset containing outdoor competitive alpine skiing scenes. It provides 10k frames of images of 6 subjects shot from 6 perspectives, of which subjects 1-5 are used for training and subject 6 is used for testing. P1 and P2 are used for model evaluation.

Implementation Details

Our experiments are conducted on one NVIDIA RTX 4090 GPU. The 2D poses used in our experiments are obtained from the pre-trained CPN (Chen et al. 2018) and ground truth. Our network parameters are optimized for 50 epochs by Adam optimizer (Kingma 2014) with an initial learning rate of 0.0002 and shrunk by 0.98 after each epoch. We consider the number of our SVTformer layer L and the hidden embedding dimension d are set to 4 and 32, respectively.

Comparison with State-of-the-art Methods

Results on Human3.6M. Table 1 compares our method with related multi-view and single-view 3D HPE methods on Human 3.6M. SVTformer outperforms all state-of-the-art single-view 3D HPE methods by a large margin. Second, our method shows competitive results with multi-view methods that require camera calibration, while our method does not depend on any pre-provided camera parameters. Moreover, our method outperforms the other multi-view methods that do not depend on camera parameters and further improves the performance using ground-truth 2D pose estimation.

Results on 3DHP. Table 2 reports the quantitative performance of STVformer with other SOTA approaches on 3DHP. Although the training set of this dataset is smaller than that of Human3.6M, it contains some outdoor scenes in addition to indoor scenes. Therefore, we train and test our

Methods	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Single-view methods																
(Zhou, Yin, and Li 2024)	44.9	46.4	42.4	44.9	48.7	40.1	44.3	55.0	58.9	47.1	48.2	42.6	36.9	48.8	40.1	46.4
(Chen et al. 2021)	41.4	43.2	40.1	42.9	46.6	51.9	41.7	42.3	53.9	60.2	45.4	41.7	46.0	31.5	32.7	44.1
(Li et al. 2023b)	39.1	42.7	38.7	40.3	44.1	50.0	41.4	38.7	53.9	61.6	43.6	40.8	42.5	29.6	30.6	42.5
(Zhao et al. 2024)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	41.6
(Zhang et al. 2022a)	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
(Ci et al. 2023)	31.7	35.4	31.7	32.3	36.4	42.4	**32.7**	31.5	**41.2**	52.7	36.5	34.0	36.2	29.5	30.2	35.6
(Peng, Zhou, and Mok 2024)	**30.1**	**32.1**	**29.1**	**30.6**	**35.4**	**39.3**	32.8	**30.9**	43.1	**45.5**	**34.7**	**33.2**	**32.7**	**22.1**	**23.0**	**33.0**
Multi-view methods (camera parameters are **given**)																
(Luvizon, Picard, and Tabia 2022)(+)	31.0	33.0	41.0	34.0	41.0	37.0	37.0	51.0	56.0	43.0	44.0	37.0	33.0	42.0	32.0	39.0
(Bultmann and Behnke 2021)	27.1	29.9	27.0	26.5	31.3	28.9	27.1	29.8	36.5	36.0	30.8	29.3	29.7	27.3	26.3	29.8
(Bartol et al. 2022)	27.5	28.4	29.3	27.5	30.1	28.1	27.9	30.8	32.9	32.5	30.8	29.4	28.5	30.5	30.1	29.1
(He et al. 2020)	25.7	27.7	23.7	24.8	26.9	31.4	24.9	26.5	28.8	31.7	28.2	26.4	23.6	28.3	23.5	26.9
(Qiu et al. 2019)(+)	24.0	26.7	23.2	24.3	24.8	22.8	24.1	28.6	32.1	**26.9**	31.0	25.6	25.0	28.1	24.4	26.2
(Iskakov et al. 2019)	**19.9**	**20.0**	**18.9**	**18.5**	**20.5**	**19.4**	**18.4**	**22.1**	**22.5**	28.7	**21.2**	**20.8**	**19.7**	**22.1**	**20.2**	**20.8**
Multi-view methods (camera parameters are **not given**)																
(Luvizon, Picard, and Tabia 2022)(+)	40.0	36.0	44.0	39.0	44.0	42.0	41.0	66.0	70.0	46.0	49.0	43.0	34.0	46.0	34.0	45.0
(Huang et al. 2020)	26.8	32.0	25.6	52.1	33.3	42.3	25.8	25.9	40.5	76.6	39.1	54.5	35.9	25.1	24.2	37.5
(Iskakov et al. 2019)	27.6	30.3	29.0	29.4	33.1	36.5	27.4	34.8	39.1	54.0	34.4	30.7	36.2	26.2	28.4	33.1
(Remelli et al. 2020)	27.3	32.1	25.0	26.5	29.3	35.4	28.8	31.6	36.4	**31.7**	31.2	29.9	**26.9**	33.7	30.4	30.2
(Zhang et al. 2024)	26.5	28.3	**23.0**	25.9	27.2	31.0	25.4	27.2	28.6	33.8	28.6	25.6	30.1	27.1	26.5	27.6
(Zhou et al. 2023)	24.8	27.7	24.3	24.9	27.7	29.8	24.5	**25.3**	30.5	33.4	28.2	24.0	28.4	24.7	24.3	26.8
Ours (CPN, T=27)	**24.5**	**27.5**	23.2	**24.4**	**25.8**	**28.7**	**23.8**	26.4	**30.0**	32.7	**26.0**	**23.9**	27.5	**22.8**	**23.2**	**26.0**
(Gordon et al. 2022)(GT, T=27)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	22.9
(Shuai, Wu, and Liu 2022)(GT, T=27)	15.5	17.1	13.7	15.5	14.0	16.2	15.8	16.5	15.8	16.1	14.5	14.5	16.9	14.3	13.7	15.3
(Zhang et al. 2024)(GT, T=27)	11.7	13.0	**10.1**	**12.1**	10.7	13.0	**12.1**	**10.7**	10.8	11.9	11.0	**11.6**	12.8	11.1	12.0	11.7
Ours (GT, T=27)	**11.6**	**12.3**	11.4	12.2	**10.6**	**12.1**	12.5	11.7	**10.3**	**10.7**	**10.7**	12.1	**10.8**	**10.9**	**11.4**	**11.4**

Table 1: Comparison with state-of-the-art 3D HPE methods on the Human3.6M dataset with P1 (mm), using CPN as 2D pose detector or GT for ground-truth 2D pose. Our results are given for temporal receptive fields below 27. (+) indicates the use of additional data. The best results are in bold.

Methods	PCK ↑	AUC ↑	P1 (mm) ↓	P2 (mm) ↓
(Kocabas, Karagoz, and Akbas 2019)	77.5	-	109.0	-
(Gholami et al. 2022)	-	-	101.5	76.5
(Wandt et al. 2021)	77.0	-	104.0	70.3
(Zhou et al. 2023)	98.7	86.8	18.0	13.1
(Zhang et al. 2024)	98.7	90.2	16.9	12.1
Ours	**99.9**	**91.6**	**12.0**	**9.1**

Table 2: Comparison results on MPI-INF-3DHP dataset.

Methods	P1 (mm)	P2 (mm)
(Wandt et al. 2021)	128.1	89.6
(Rhodin et al. 2018)	85.0	-
(Gordon et al. 2022)	65.5	-
(Zhang et al. 2024)	63.2	48.5
(Zhou et al. 2023)	62.6	49.4
Ours	**59.9**	**47.6**

Table 3: Comparisons on Ski-Pose dataset.

model end-to-end on the 3DHP dataset. The results depict that our method achieves the best on all four evaluation metrics. This emphasizes that our method performs well on both indoor and outdoor datasets.

Results on Ski-Pose. Table 3 shows the results of our method trained and tested from scratch on the Ski-Pose dataset. The results of our method surpass all other methods, demonstrating the superiority of our method on more challenging outdoor scene datasets.

Qualitative Result. We also provide a visualization comparing our SVTformer, the method of (Zhang et al. 2024), and ground truth in Fig. 4. It can be observed that our model

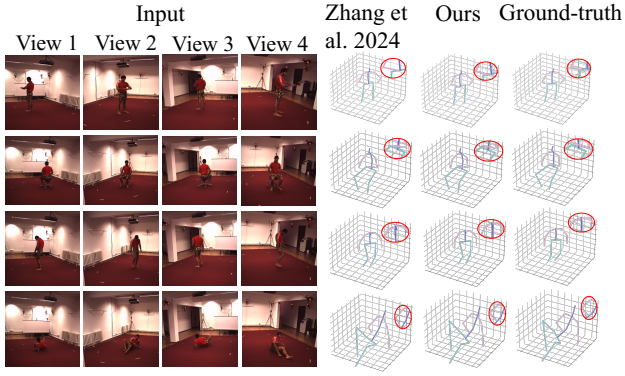

Figure 4: Qualitative results on Human3.6M.

can get closer estimates to the ground truth for both easy and difficult poses, especially those with self-occlusion. We also visualize the attention maps of the SVT encoder and the attended SVT PE, shown in Fig. 2. It can be seen that the attention maps of the attended SVT PE focus more on local joints, viewpoints, and temporal features. In contrast, the attention maps of the SVT encoder become increasingly sparse as the layers deepen, which is beneficial for capturing global relationships.

Ablation Study

Effect of the variants of SVT Transformer. We first compare the performance of four transformer models for learning spatial, view, and temporal relationships. As shown in Table 4, the sequentially alternating spatial-view-temporal transformer (our SVTformer) performs best, outperforming

Model	P1 (mm)	P2 (mm)
Sequentially alternating model (SVTformer)	**26.0**	**19.5**
Parallel weighted model	28.7	22.1
Joint sequential and parallel model	26.5	19.9
Joint sequential and parallel model*	27.9	20.6

Table 4: Effect of the variants of SVT transformer.

Sequential correspondence orders	P1(mm)	P2(mm)
$\mathbb{V} \rightarrow \mathbb{S} \rightarrow \mathbb{T}$	29.1	22.1
$\mathbb{V} \rightarrow \mathbb{T} \rightarrow \mathbb{S}$	28.8	21.9
$\mathbb{T} \rightarrow \mathbb{V} \rightarrow \mathbb{S}$	28.3	21.1
$\mathbb{T} \rightarrow \mathbb{S} \rightarrow \mathbb{V}$	28.0	21.4
$\mathbb{S} \rightarrow \mathbb{T} \rightarrow \mathbb{V}$	26.2	19.8
$\mathbb{S} \rightarrow \mathbb{V} \rightarrow \mathbb{T}$	**26.0**	**19.5**

Table 5: Effect of the sequential correspondence orders in attended SVT PE and SVT encoder. $\mathbb{S}$ is attended spatial PE and spatial attention. $\mathbb{V}$ is attended view PE and view attention. $\mathbb{T}$ is attended temporal PE and temporal attention.

the parallel structure. The parallel weighted model considers the correlation from three aspects in parallel. Since each branch only learns one aspect of the association, the adaptive fusion part aggregates all the associations at the end, resulting in insufficient learning of the three associations of the joints, view, and time. This result verifies the effectiveness of the sequential spatial-view-temporal attention modules introduced in our SVTformer that continuously reshapes 2D features and sequentially mines the spatial, view, and temporal relationships of 2D poses. Two joint sequential and parallel models perform better than the parallel weighted model, suggesting that sequential modeling can make up for the insufficient learning of parallel relationship modeling and further improve the performance of parallel models. It reveals that for three aspects of correlation learning in multi-view HPE, their modeling methods are also very important.

Effect of the sequential correspondence orders. Table 5 shows how the learning order of temporal, spatial, and multi-view correspondences in patch embeddings and attention modules affects the model effect in our SVTformer. From the experimental results, we can see that the best performance is achieved by first modeling the spatial relationship between joints of each view, then considering the relationship between joints of different views, and finally learning the temporal motion of joints of each view. On the contrary, the methods that first learn cross-view or temporal features perform worse. This is because the pose information of our model is reshaped by continuously merging and decoupling 2D pose features and then sequentially interacting with each view spatial feature, cross-view spatial feature interaction, and each view temporal interaction to mine the features of joints, which gradually makes the original 2D pose features more meaningful.

Effect of each component. To diagnose the role of each module in SVTformer, we removed attended SVT PE and SVT encoders, as well as their submodules, as shown in Table 6. The results show that the performance of the model

Attended SVT PE			SVT Encoder			P1 (mm)	P2 (mm)
Spatial	View	Temporal	Spatial	View	Temporal		
×	✓	✓	✓	✓	✓	26.8	20.6
✓	×	✓	✓	✓	✓	26.6	19.9
✓	✓	×	✓	✓	✓	26.1	19.9
✓	✓	✓	×	✓	✓	27.2	20.7
✓	✓	✓	✓	×	✓	26.5	20.1
✓	✓	✓	✓	✓	×	26.7	20.3
×	×	×	✓	✓	✓	26.5	20.1
✓	✓	✓	×	×	×	27.5	20.5
✓	✓	✓	✓	✓	✓	**26.0**	**19.5**

Table 6: Effect of each component in SVTformer.

View number	1	2	3	4
P1 (mm)	35.2	30.6	29.1	**26.0**
P2 (mm)	26.4	24.2	22.4	**19.5**

Table 7: Effect of the number of fused views.

suffers a drop when any module is removed, indicating that both the patch embeddings and encoder based on SVT attention are beneficial to the extraction of multi-view 2D pose information. In addition, comparing the two modules of attended SVT PE and SVT encoder, we can see that attended SVT PE contributes more, giving the model a performance improvement of 1.5mm and 1mm on P1 and P2, respectively, indicating that SVT attention is needed to further enhance representation learning based on the SVT-enhanced embeddings. In particular, we find that the spatial attention modules in both two modules play a more significant role than the cross-view and temporal attention modules, indicating that the joint-to-joint spatial features of each view can mine more robust multi-view pose representations.

Effect of the number of fused views. We compare the use of different numbers of view fusion for 3D HPE, and the results are shown in Table 7. As the number of views increases, the performance of our model steadily increases, and it performs best when it reaches 4 views. This shows that our sequential modeling of space, view, and time can effectively fuse information from multiple views and compensate well for the missing joint information in a single view.

Conclusion

In this paper, we propose SVTformer, a new transformer-based approach for multi-view 3D HPE, by separately exploring the spatial-view-temporal correlation in a local-to-global manner. Based on the transformer, we introduce attended patch embedding to capture the local relationships within each part of spatial, view, and temporal inputs and devise the stacked spatial-view-temporal transformer encoder to obtain the global association perception. Extensive experimental results show that our model achieves state-of-the-art performance on three 3D HPE benchmarks.

Acknowledgments

This work was jointly supported by the National Natural Science Foundation of China (No.62373009), the Nat-

ural Science Foundation of Guangdong Province (No. 2024A1515012089), and the Shenzhen Innovation in Science and Technology Foundation for The Excellent Youth Scholars (No. RCYX20231211090248064).

References

Bartol, K.; Bojanić, D.; Petković, T.; and Pribanić, T. 2022. Generalizable human pose triangulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11028–11037.

Bultmann, S.; and Behnke, S. 2021. Real-time multi-view 3D human pose estimation using semantic feedback to smart edge sensors. *Robotics: Science and Systems*.

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.

Chu, H.; Lee, J.-H.; Lee, Y.-C.; Hsu, C.-H.; Li, J.-D.; and Chen, C.-S. 2021. Part-aware measurement for robust multi-view multi-human 3d pose estimation and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1472–1481.

Ci, H.; Wu, M.; Zhu, W.; Ma, X.; Dong, H.; Zhong, F.; and Wang, Y. 2023. Gfpose: Learning 3d human pose prior with gradient fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4800–4810.

Fasel, B.; Spörri, J.; Gilgien, M.; Boffi, G.; Chardonnens, J.; Müller, E.; and Aminian, K. 2016. Three-dimensional body and centre of mass kinematics in alpine ski racing using differential GNSS and inertial sensors. *Remote Sensing*, 8(8): 671.

Gholami, M.; Rezaei, A.; Rhodin, H.; Ward, R.; and Wang, Z. J. 2022. Self-supervised 3D human pose estimation from video. *Neurocomputing*, 488: 97–106.

Gordon, B.; Raab, S.; Azov, G.; Giryes, R.; and Cohen-Or, D. 2022. FLEX: extrinsic parameters-free multi-view 3D human motion reconstruction. In *European Conference on Computer Vision*, 176–196. Springer.

He, Y.; Yan, R.; Fragkiadaki, K.; and Yu, S.-I. 2020. Epipolar transformer for multi-view human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 1036–1037.

Hua, G.; Liu, H.; Li, W.; Zhang, Q.; Ding, R.; and Xu, X. 2022. Weakly-supervised 3D human pose estimation with cross-view U-shaped graph convolutional network. *IEEE Transactions on Multimedia*, 25: 1832–1843.

Huang, F.; Zeng, A.; Liu, M.; Lai, Q.; and Xu, Q. 2020. DeepFuse: An IMU-aware network for real-time 3D human pose estimation from multi-view image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 429–438.

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.

Iskakov, K.; Burkov, E.; Lempitsky, V.; and Malkov, Y. 2019. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7718–7727.

Jiang, B.; Hu, L.; and Xia, S. 2023. Probabilistic triangulation for uncalibrated multi-view 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14850–14860.

Kadkhodamohammadi, A.; and Padoy, N. 2021. A generalizable approach for multi-view 3d human pose regression. *Machine Vision and Applications*, 32(1): 6.

Kim, H.-W.; Lee, G.-H.; Nam, W.-J.; Jin, K.-M.; Kang, T.-K.; Yang, G.-J.; and Lee, S.-W. 2024. MHCanonNet: Multi-Hypothesis Canonical lifting Network for self-supervised 3D human pose estimation in the wild video. *Pattern Recognition*, 145: 109908.

Kim, H.-W.; Lee, G.-H.; Oh, M.-S.; and Lee, S.-W. 2022. Cross-view self-fusion for self-supervised 3d human pose estimation in the wild. In *Proceedings of the Asian Conference on Computer Vision*, 1385–1402.

Kingma, D. 2014. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kocabas, M.; Karagoz, S.; and Akbas, E. 2019. Self-supervised learning of 3d human pose using multi-view geometry. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1077–1086.

Li, H.; Shi, B.; Dai, W.; Zheng, H.; Wang, B.; Sun, Y.; Guo, M.; Li, C.; Zou, J.; and Xiong, H. 2023a. Pose-oriented transformer with uncertainty-guided refinement for 2d-to-3d human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1296–1304.

Li, W.; Liu, H.; Tang, H.; and Wang, P. 2023b. Multi-hypothesis representation learning for transformer-based 3D human pose estimation. *Pattern Recognition*, 141: 109631.

Li, W.; Liu, H.; Tang, H.; Wang, P.; and Van Gool, L. 2022. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13147–13156.

Li, W.; Liu, M.; Liu, H.; Wang, P.; Cai, J.; and Sebe, N. 2024. Hourglass Tokenizer for Efficient Transformer-Based 3D Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 604–613.

Luvizon, D. C.; Picard, D.; and Tabia, H. 2022. Consensus-based optimization for 3D human pose estimation in camera coordinates. *International Journal of Computer Vision*, 130(3): 869–882.

Ma, H.; Chen, L.; Kong, D.; Wang, Z.; Liu, X.; Tang, H.; Yan, X.; Xie, Y.; Lin, S.-Y.; and Xie, X. 2021. Transfusion: Cross-view fusion with transformer for 3d human pose estimation.

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.

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.

Qiu, H.; Wang, C.; Wang, J.; Wang, N.; and Zeng, W. 2019. Cross view fusion for 3d human pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4342–4351.

Rajasegaran, J.; Pavlakos, G.; Kanazawa, A.; Feichtenhofer, C.; and Malik, J. 2023. On the benefits of 3d pose and tracking for human action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 640–649.

Remelli, E.; Han, S.; Honari, S.; Fua, P.; and Wang, R. 2020. Lightweight multi-view 3D pose estimation through camera-disentangled representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6040–6049.

Rhodin, H.; Spörri, J.; Katircioglu, I.; Constantin, V.; Meyer, F.; Müller, E.; Salzmann, M.; and Fua, P. 2018. Learning monocular 3d human pose estimation from multi-view images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8437–8446.

Shuai, H.; Wu, L.; and Liu, Q. 2022. Adaptive multi-view and temporal fusing transformer for 3d human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4122–4135.

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.

Wan, X.; Chen, Z.; and Zhao, X. 2023. View consistency aware holistic triangulation for 3D human pose estimation. *Computer Vision and Image Understanding*, 236: 103830.

Wandt, B.; Rudolph, M.; Zell, P.; Rhodin, H.; and Rosenhahn, B. 2021. Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13294–13304.

Wang, H.-K.; Huang, M.; Zhang, Y.; and Song, K. 2023a. Multi-View 3D Human Pose and Shape Estimation with Epipolar Geometry and Mix-Graphormer. In *2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP)*, 28–32. IEEE.

Wang, T.; Liu, H.; Song, P.; Guo, T.; and Shi, W. 2022. Pose-guided feature disentangling for occluded person re-identification based on transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 2540–2549.

Wang, X.; Fang, Z.; Li, X.; Li, X.; Chen, C.; and Liu, M. 2024. Skeleton-in-context: Unified skeleton sequence modeling with in-context learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2436–2446.

Wang, X.; Zhang, W.; Wang, C.; Gao, Y.; and Liu, M. 2023b. Dynamic Dense Graph Convolutional Network for Skeleton-based Human Motion Prediction. *IEEE Transactions on Image Processing (TIP)*.

Yu, B. X.; Zhang, Z.; Liu, Y.; Zhong, S.-h.; Liu, Y.; and Chen, C. W. 2023. Gla-gcn: Global-local adaptive graph convolutional network for 3d human pose estimation from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8818–8829.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14985–14995.

Zhang, J.; Tu, Z.; Yang, J.; Chen, Y.; and Yuan, J. 2022a. 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.

Zhang, L.; Zhou, K.; Lu, F.; Zhou, X.-D.; and Shi, Y. 2024. Deep Semantic Graph Transformer for Multi-View 3D Human Pose Estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7205–7214.

Zhang, Y.; Wang, C.; Wang, X.; Liu, W.; and Zeng, W. 2022b. Voxeltrack: Multi-person 3d human pose estimation and tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 2613–2626.

Zhao, L.; Peng, X.; Tian, Y.; Kapadia, M.; and Metaxas, D. N. 2019. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3425–3435.

Zhao, Q.; Zheng, C.; Liu, M.; and Chen, C. 2024. A single 2d pose with context is worth hundreds for 3d human pose estimation. *Advances in Neural Information Processing Systems*, 36.

Zhou, F.; Yin, J.; and Li, P. 2024. Lifting by image–leveraging image cues for accurate 3d human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7632–7640.

Zhou, K.; Zhang, L.; Lu, F.; Zhou, X.-D.; and Shi, Y. 2023. Efficient Hierarchical Multi-view Fusion Transformer for 3D Human Pose Estimation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7512–7520.