Computer Vision and Image Understanding 243 (2024) 103992

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

SlowFastFormer for 3D human pose estimation

Lu Zhou [a], Yingying Chen [a,*], Jinqiao Wang [a,b,c,d]

[a] Foundation Model Research Center, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[b] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
[c] Wuhan AI Research, Wuhan 430073, China
[d] Peng Cheng Laboratory, Shenzhen 518066, China

ARTICLE INFO

Communicated by Si Liu

Keywords:
SlowFastFormer
Transformer
Blending
3D human pose estimation
Hierarchical supervision

ABSTRACT

3D human pose estimation in videos aims at locating the human joints in the 3D space given a temporal sequence. Motion information and skeleton context are two significant elements for pose estimation in videos. In this paper, we propose a SlowFastFormer (slow-fast transformer) network where two branches with different input rates are composed to encode these two different kinds of context. For the slow branch, skeleton context is well learned at a higher frame rate. For the fast branch, motion information is captured at a lower frame rate. Through these two branches, different kinds of context are encoded separately. We fuse these two branches at a later stage to fully utilize the skeleton context and motion information. Afterwards, a blending module is developed to promote the message exchange among multiple branches. In the blending stage, different kinds of context information are exchanged and feature representation is enhanced consequently. Lastly, a hierarchical supervision scheme is tailored where predictions of different levels are inferred in a progressive manner. Our approach achieves competitive performance with lower computation complexity on several benchmarks, i.e., Human3.6M, MPI-INF-3DHP and HumanEva-I.

1. Introduction

3D human pose estimation serves as a fundamental task in the community of computer vision. It has been widely applied in the field of human–robot interactions (Garcia-Salguero et al., 2019; Gui et al., 2018), action recognition (Anon, 2023b; Gedamu et al., 2023; Wu et al., 2023; Peng et al., 2021), etc. For video task, 3D human pose estimation mainly locates the joints of the target frame given a video sequence. Among these methods, 2D-3D lifting approaches achieve much better performance than one-stage counterparts. These approaches firstly predict the 2D human joint coordinates and transform the 2D results into 3D predictions afterwards. Despite the progress achieved, the task is still challenged by the ill-posed problem caused by the depth ambiguity.

To conquer this dilemma, recent approaches employ transformer structures to learn the joint relations or temporal relations. PoseFormer (Zheng et al., 2021) firstly employs the transformer module to construct both the joint relations and temporal relations. StridFormer (Li et al., 2022a) employs the transformer module and predicts the human poses in a progressive manner. MHFormer (Li et al., 2022b) proposes a Multi-Hypothesis Transformer network where interactions among multiple hypotheses are involved. Nevertheless, these works neglect the significance of both the motion information and skeleton context which provide explicit cues for the task of pose estimation in videos.

Though the 3D human pose estimation in videos is a regression task, it shares some common characteristics with the action classification. For action recognition, network relies on the spatial semantics and temporal information to infer final action label. For pose estimation, regression of the target frame relies on the motion change of adjacent frames as well except for the skeleton context. To learn both the spatial semantics and temporal context for action inference, SlowFast (Feichtenhofer et al., 2019) devises a dual path framework where the fast path mainly learns the temporal information and the slow path mainly learns the spatial semantics. Inspired by this, in this paper, we propose a slow-fast transformer network to learn the fine motion detail and anatomy information simultaneously via transformer module. Besides, the new framework with lower computational complexity is more efficient than previous approaches.

The whole framework is composed of two stages: parallel encoding stage and blending stage. For the parallel encoding stage, two branches with different input rates are exploited to encode the motion information and skeleton context separately. For the slow refreshing branch, input rates and temporal resolution are lower. The branch focuses more on the skeleton context of input sequence. For the fast refreshing branch, input rates and temporal resolution are set $\alpha\times$ higher than the slow path. The branch focuses more on the fine motion change which poses great significance for some fast changing motion actions. Different from the two-path way design in Feichtenhofer et al. (2019),

* Corresponding author at: Foundation Model Research Center, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.
E-mail address: yingying.chen@nlpr.ia.ac.cn (Y. Chen).

https://doi.org/10.1016/j.cviu.2024.103992
Received 20 August 2023; Received in revised form 16 November 2023; Accepted 11 March 2024
Available online 24 March 2024

channel number of fast branch is kept the same with the slow branch. The performance drops drastically when channel number of fast branch is cut in half. Similarly, we bridge the two branches with the bilateral connections. Eventually, we fuse the slow and fast branch to integrate the different kinds of context and the new formulated branch is dubbed as uniform branch.

Interactions among various branches pose great significance for multi-branch framework. Different levels of features are strengthened though the multi-branch message passing, which benefits the final prediction. Motivated by this, a blending module is proposed to facilitate the message passing among slow, fast and uniform branches. The blending module takes all the three branches as input and employs the transformer structure to build the cross-branch relations. We split the features of each branch into M subgroups, where M represents the number of human joints. The interaction within this stage mainly concentrates on the spatial relations. From other point of view, these three branches can also be regarded as three different hypotheses and interaction process can enhance the representation of each branch. An MLP (Multi-Layer Perceptron) is utilized to fuse the three different branches finally and output the final predictions.

The parallel encoding stage and the blending stage enhance the feature representation progressively. Following the structure design of the main network, we design a hierarchical supervision scheme to refine the predicted poses progressively. Firstly, we enforce full supervision after the parallel encoding stage where supervision of each frame is provided. Ground-truth joint coordinates of k frames and αk frames is supervised for slow path and fast path respectively. Specifically, the intermediate predictions make the parallel encoding speed-aware. In addition, we add a consistency regulation item to achieve the prediction consistency between two pose sets. For the blending stage where features of k frames are fused initially, only supervision of central frame is provided. The supervision is enforced on each joint separately and makes the blending joint-aware. These two supervisions play different roles and reduce the location error progressively.

To verify the effectiveness, we conduct the experiments on several benchmarks, i.e., Human3.6M (Ionescu et al., 2013), MPI-INF-3DHP (Mehta et al., 2017a) and HumanEva-I (Sigal et al., 2010). Competitive results are achieved on these datasets. Compared with the previous 3D human pose estimation approaches, it is the first attempt to explore the effectiveness of slow-fast architecture in 3D human pose estimation, i.e., a regression task. However, the work differs a lot from the prime slow-fast network in video understanding: (1) The feature enhancement procedure is a progressive paradigm which is quite different from Feichtenhofer et al. (2019). (2) We formulate the whole framework as a three-branch structure and assume the multi-branch representations as different hypotheses which differs from Feichtenhofer et al. (2019). (3) We employ different kinds of transformer to conduct the feature modelling which is more effective. Contributions of the paper can be summarized as follows:

- We propose a SlowFastFormer network which consists of parallel encoding and blending stages to enhance the features progressively with lower computational complexity.
- We propose a parallel encoding module which is composed of slow and fast branches to encode both the skeleton context and motion information in parallel.
- We propose a blending module to promote the message passing between three different branches, i.e., slow branch, fast branch and uniform branch. Interactions facilitate the cross-branch communication and strengthen the multi-level features further.
- We propose a hierarchical supervision scheme to refine the predictions progressively. Whole learning process is eased under this schedule.

2. Related works

2.1. 3D human pose estimation

3D human pose estimation from monocular images predicts the 3D joint coordinates given only one-view input. Approaches in this field are divided into two categories, i.e., one-stage approach and 2D-3D lifting approach. One-stage approaches (Liu et al., 2019; Pavlakos et al., 2017; Han et al., 2022; Zhao et al., 2019) mainly predict the human joints in a single shot without the participation of intermediate 2D predictions. Nonetheless, these approaches usually demand higher computations and complex structure designs which are seldom employed. In contrast, lifting approaches are divided into two stages, i.e., 2D prediction stage and 2D-3D projection stage. For the 2D prediction stage, model (Sun et al., 2019; Jiang et al., 2023; Tian et al., 2021) trained on large scale 2D human pose estimation benchmarks is employed to predict the intermediate 2D joints. For the second stage, the 2D-3D mapping relation is well learned and most approaches are devoted to improving the projection precision. Among these approaches, there exist some works which only predict the 3D predictions based on single frame input. SimpleBaseline (Martinez et al., 2017) devises a fully connected residual network and achieves competitive results over the projection task. Grammar3D (Fang et al., 2018) employs different types of pose grammar to learn the inherent human constraints. VIPose (Wei et al., 2019) proposes a view invariant network to address the issue of view diversity. GraphHourglass (Xu and Takano, 2021) takes advantage of stacked graph hourglass module to encode both the local and global information. HTNet (Cai et al., 2023) proposes a human topology aware network to build the context of joint, part and body. GridConv (Kang et al., 2023) conquers the pose estimation via grid convolution. Different from the approaches mentioned above, inferring poses from video sequence acquires more attention due to the merits brought by the additional temporal dimension. Video3D (Pavllo et al., 2019) employs TCN to aggregate the information of multiple frames and assist the prediction of central frame. Motivated by this, works (Liu et al., 2020b; Chen et al., 2021) devote great efforts to the improvement of TCN and achieve leading performance. Works (Hu et al., 2021; Wang et al., 2020) exploit spatial–temporal graph convolutions to build the spatial and temporal relations of the input sequence. Transformer-based approach (Zheng et al., 2021) exploits vanilla transformer to construct the spatial and temporal relations of the input skeleton sequence. StrideFormer (Li et al., 2022a) utilizes the stride convolution to aggregate the local contexts. MHFormer (Li et al., 2022b) constructs strong relations among different hypotheses to remove the prediction ambiguity. P-STMO (Shan et al., 2022) introduces a pretraining scheme in this task and finetunes afterwards. PoseFormerV2 (Zhao et al., 2023) fuses the features from both the time domain and frequency domain. HDFormer (Chen et al., 2023) proposes a multi-order attention module to promote the interactions of different orders. STCFormer (Tang et al., 2023) models the spatial and temporal information in parallel via Spatio-Temporal Criss-cross attention. UPS (Foo et al., 2023) unifies different human understand tasks and achieves promising results on 3D pose estimation. Different from the mentioned approaches, our approach makes use of the slow-fast transformer structure to enhance the feature representation in a progressive manner. Various contexts are disclosed under the proposed formulation. It is the first attempt to explore the effectiveness of slow-fast paradigm. Besides, we propose a blending module to promote the message passing between different branches which can be regarded as different hypotheses. Different from MHFormer, detailed context is embedded in different branches and joint relations are explicitly built in the blending module.

2.2. Vision transformer

Transformer (Vaswani et al., 2017) with powerful self-attention mechanism is widely applied in NLP and becomes prevalent in computer vision due to its excellent performance. ViT (Dosovitskiy et al.,

2020) is a widely used vision transformer for image classification task. After that, a series of works devote to the improvements of prime ViT. TNT (Han et al., 2021) and T2TViT (Yuan et al., 2021a) make great efforts to the improvement of tokenization. SiT (Zong et al., 2022) can softly aggregate redundant tokens instead of dropping them hardly. ATS (Fayyaz et al., 2022) dynamically samples the informative tokens and the token number is adaptive to different inputs. CVPT (Chu et al., 2021) develops a dynamic position encoding module to adapt to different input resolutions. Swin-Transformer (Liu et al., 2021) builds a hierarchical transformer which can model at different scales. PVT (Wang et al., 2021) proposes a Pyramid Vision Transformer network which combines both the merits of CNN and transformer and operates at various scales as well. EdgeViTs (Pan et al., 2022) designs effective ViT backbones adapting to the edge devices and achieves competitive performance compared with light-weight CNNs. PTQ4ViT (Yuan et al., 2022) studies the quantization problem of ViT and achieves lower performance drop at 8-bit quantization. CETNet (Wang et al., 2022) explores the effectiveness of convolution embedding in transformer-based framework and serves as a common backbone for vision task. In addition to the image classification, tasks such as object detection (Carion et al., 2020; Li et al., 2022c), pose estimation (Li et al., 2021; Yuan et al., 2021b), segmentation (Yuan et al., 2021b), captioning (Anon, 2023a) also make use of vision transformer which greatly improves the performance. Inspired by this, we take transformer block as the basic structure and devise a brand framework to refine the features asymptotically. We combine the vanilla transformer block and cross transformer block to perform the refinement progressively.

2.3. SlowFast network

SlowFast (Feichtenhofer et al., 2019) devises a dual path framework where temporal information and spatial semantics are learned for action recognition. The network achieves outstanding performance and the architecture is employed by following works. A dual attention SlowFast network (Wei et al., 2022) is designed for video action recognition, where a cross-modality dual attention fusion module is proposed to exchange spatial–temporal information. A SlowFast network for audio recognition is proposed in Kazakos et al. (2021) where separable convolutions and multi-level lateral connections are employed. A SlowFast network is exploited in Ahn et al. (2023) for sign language recognition. Different from the above approaches, our proposed method adopts transformer as the main module instead of CNNs. We divide the whole framework into two different stages and a blending module is designed to enhance features of the first stage. Besides, a hierarchical supervision scheme is employed to refine the predictions progressively.

3. Methodology

Overview of the whole framework can be found in Fig. 1. The proposed SlowFastFormer (slow-fast transformer) is split into two stages, i.e., parallel encoding stage and blending stage. For the parallel encoding part, slow and fast branches are separately encoded and communicated by bilateral connections. Fusing the slow and fast branches forms our uniform path. For the blending stage, outputs of the slow, fast and uniform branches constitute the input of the blending module. Full and single frame supervisions are enforced at the tail of parallel encoding stage and blending stage respectively.

3.1. Parallel encoding stage

Parallel encoding module is composed of two different branches to encode motion information and skeleton context respectively. For the slow branch, input of k frames is employed. For the fast branch, we adopt clip of $9k$ frames as input which is $9\times$ denser than the slow counterpart. Both of these two branches employ transformer blocks to build temporal relations of different sequences. We will give a brief introduction on the feature encoding of the mentioned two branches.

Slow branch. For the slow branch, input is denoted as $X^s = [x_1^s, x_2^s, \dots, x_k^s] \in R^{k\times M\times 2}$, where k represents the sequence length and M represents the joint number. The input can be projected into $Z^s = [x_1^s E^s, x_2^s E^s, \dots, x_k^s E^s] \in R^{k\times C}$, where E^s represents the encoder layer. Positional encoding $E_{pos}^s \in R^{k\times C}$ is added to the input projection. The input can be denoted as follows:

$$Z_0^s = Z^s + E_{pos}^s. \tag{1}$$

For the encoder part, vanilla transformer is utilized. We acquire query matrix Q, key matrix K and value matrix V as follows:

$$Q = ZW_Q, K = ZW_K, V = ZW_V, \tag{2}$$

where $W_Q \in R^{k\times C}, W_K \in R^{k\times C}, W_V \in R^{k\times C}$ are the projection matrices. Attention is calculated as follows:

$$Atten(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{C}}). \tag{3}$$

As a common formulation, multi-head self-attention (MSA) is applied.

$$\begin{aligned} &MSA(Q, K, V) = Cat(H_1, H_2, H_3, \dots, H_h)W_{out}, \\ &where \quad H_i = Atten(Q_i, K_i, V_i), i = 1, 2, 3..., h, \end{aligned} \tag{4}$$

where W_{out} denotes the projection matrix and h is the head number. There are L layers in the encoder,

$$\begin{aligned} Z_l^{s'} &= MSA(LN(Z_{l-1}^s)) + Z_{l-1}^s, \\ Z_l^s &= FFN(LN(Z_l^{s'})) + Z_l^{s'}, \\ \overline{Z}^s &= LN(Z_L^s), \end{aligned} \tag{5}$$

where FFN represents the feed-forward network and LN is the layer normalization layer.

Fast branch. Fast branch adopts the same encoding procedure as slow branch. The only difference comes in the input length. Channel number is kept the same with slow branch due to the large performance degradation caused by the channel cropping. We can acquire the input of fast branch as:

$$Z_0^f = Z^f + E_{pos}^f, \tag{6}$$

where $Z_0^f \in R^{9k\times C}$ and k is the sequence length of slow branch. Encoding process can be denoted as:

$$\begin{aligned} Z_l^{f'} &= MSA(LN(Z_{l-1}^f)) + Z_{l-1}^f, \\ Z_l^f &= FFN(LN(Z_l^{f'})) + Z_l^{f'}, \\ \overline{Z}^f &= LN(Z_L^f), \end{aligned} \tag{7}$$

where $\overline{Z}^f$ represents the output of fast branch.

Lateral connection. Lateral connection is constructed to alleviate the context gap of dual branches. Herein, we simply utilize the bilateral addition operation where no additional convolutions are involved. We find that additional convolution operation may even cause the performance degradation.

For the forward direction (slow to fast), we upsample the features of the slow branch at first and then add it with the feature of fast branch:

$$Z_l^f = Z_l^f + Up(Z_l^s), \tag{8}$$

where Up represents the upsampling operation and frame number of $Up(Z_l^s)$ is $9k$.

For the backward direction (fast to slow), we downsample the features of the fast branch and then add it with the feature of slow branch:

$$Z_l^s = Z_l^s + Down(Z_l^f), \tag{9}$$

where $Down$ represents the downsampling operation and frame number of $Down(Z_l^f)$ is k.

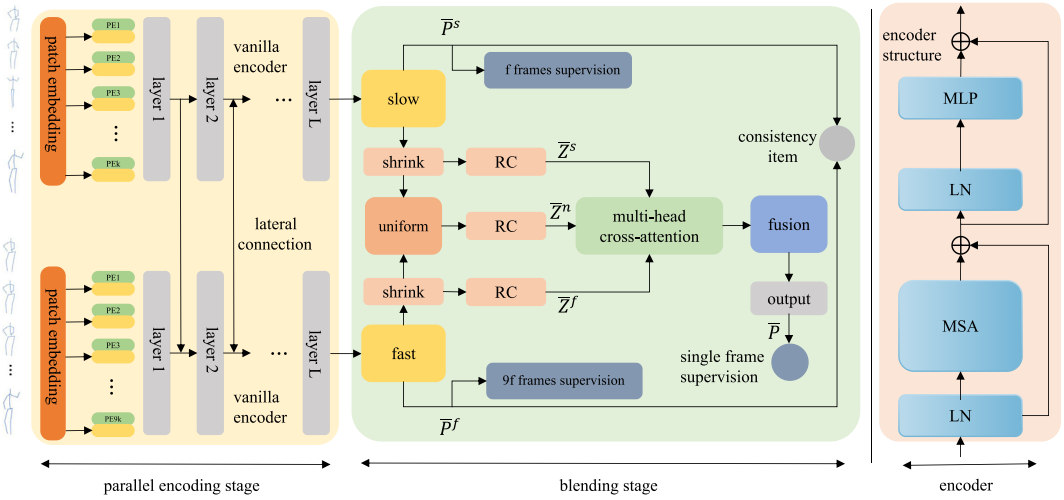

Fig. 1. Structure of the proposed SlowFastFormer network for 3D human pose estimation. Skeleton sequences with different sampling rates constitute the input. Whole framework consists of parallel encoding stage and blending stage. Hierarchical supervision scheme is enforced at the end of each stage. Operation 'shrink' changes the frame number to 1 and 'RC' expands the channels of each branch.

3.2. Blending stage

Information interaction poses great significance for multi-branch frameworks. Though the message passing, features of different paths can be enhanced by the other branches. In this paper, we design a blending module which mainly facilitates message passing among different branches.

Uniform branch. Uniform branch is newly formed which fuses the features of slow and fast branches. For this branch, it incorporates different levels of contexts. From the point of multiple hypotheses, the new branch can serve as another hypothesis which enjoys the merits of both slow and fast branches. Firstly, we reduce the channel number of the two branch features to 1 via convolution operation:

$$\begin{aligned} \overline{Z}^s &= Conv(\overline{Z}^s), \\ \overline{Z}^f &= Conv(\overline{Z}^f), \end{aligned} \tag{10}$$

where $Conv$ represents the 1×1 convolution which reduces the temporal dimension into 1. Fusion of the two branch features can be denoted as:

$$\overline{Z}^n = \overline{Z}^s + \overline{Z}^f. \tag{11}$$

Blending module. Blending module takes both the three branches as input. Firstly, channel number of the three branches is expanded via the fc layer ('RC' in Fig. 1). After that, the feature of each branch is reshaped into $R^{M \times C}$, where M represents the joint number and C represents the channel number. In this case, sequence length is the joint number and each channel represents the feature of a specific human joint. Interactions across different branches can also be regraded as the feature enhancement of each human joint.

We take advantage of transformer to conduct the interaction. In the blending transformer block (see Fig. 2), features from slow, fast and uniform branches serve as the query, key and value alternately. If feature of slow branch serves as the value, features of fast and uniform branches are designated as the query and key respectively. Formally,

$$\overline{Z}_l^{s'} = MCA^s(LN(\overline{Z}_{l-1}^s), LN(\overline{Z}_{l-1}^f), LN(\overline{Z}_{l-1}^n)) + \overline{Z}_{l-1}^s, \tag{12}$$

where MCA indicates the multi-head cross-attention (Li et al., 2022b). Features of fast and uniform $\overline{Z}_l^{f'}$, $\overline{Z}_l^{n'}$ branch can be acquired following the same way. For the MLP, we can formulate it as:

$$\begin{aligned} \overline{Z}'_l &= cat(\overline{Z}_l^{s'}, \overline{Z}_l^{f'}, \overline{Z}_l^{n'}), \\ cat(\overline{Z}_l^s, \overline{Z}_l^f, \overline{Z}_l^n) &= FFN(LN(\overline{Z}'_l)) + \overline{Z}'_l. \end{aligned} \tag{13}$$

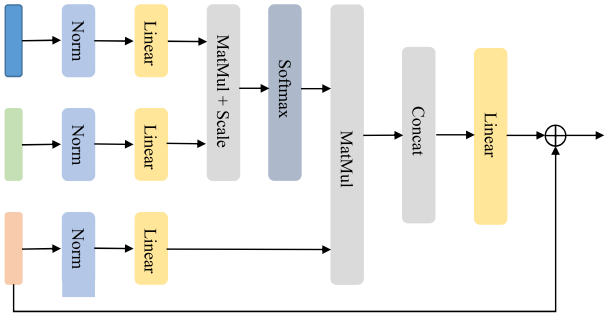

Fig. 2. Structure of the multi-head cross-attention (MCA).

where cat represents the concatenation operation. Output of the blending stage can be denoted as:

$$\begin{aligned} \overline{Y}^s &= LN(\overline{Z}_L^s), \\ \overline{Y}^f &= LN(\overline{Z}_L^f), \\ \overline{Y}^n &= LN(\overline{Z}_L^n). \end{aligned} \tag{14}$$

Afterwards, we sum features of each branch together and predict the final pose based on it:

$$\overline{Y} = \overline{Y}^s + \overline{Y}^f + \overline{Y}^n. \tag{15}$$

Through the blending stage, interactions among different branches are performed. Compared with the first stage where temporal modelling is performed, blending stage mainly focuses on the human context interactions among different branches. The differences between ours and Cross-Hypothesis Interaction (CHI) proposed in Li et al. (2022b) are two-fold: (i) Interaction elements in the blending stage encode different levels of context which are more explainable. (ii) Temporal dimension in CHI is replaced with joint dimension in the blending module where expertise has been changed. The computations are effectively reduced in this case.

3.3. Hierarchical supervision

Features from parallel encoding stage represent different refreshing speeds of input sequence. To make the feature be aware of the different refreshing speeds, we enforce full supervisions on both the branches. For the slow branch, k frames supervision is imposed. For

Table 1
Quantitative results of 3D HPE approaches and 'Ours' denotes the SlowFastFormer proposed in the paper. MPJPE under Protocol 1 serves as the evaluation metric.

Protocol 1		Dir.	Disc.	Eat.	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Average
Zhao et al. (2019)	CVPR'19	48.2	60.8	51.8	64.0	64.6	53.6	51.1	67.4	88.7	57.7	73.2	65.6	48.9	64.8	51.9	60.8
Liu et al. (2020a)	ECCV'20	46.3	52.2	47.3	50.7	55.5	67.1	49.2	46.0	60.4	71.1	51.5	50.1	54.5	40.3	43.7	52.4
Sharma et al. (2019)	ICCV'19	48.6	54.5	54.2	55.7	62.2	72.0	50.5	54.3	70.0	78.3	58.1	55.4	61.4	45.2	49.7	58.0
Xu and Takano (2021)	CVPR'21	45.2	49.9	47.5	50.9	54.9	66.1	48.5	46.3	59.7	71.5	51.4	48.6	53.9	39.9	44.1	51.9
Pavllo et al. (2019) ($k = 243$)*	CVPR'19	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
Lin and Lee (2019) ($k = 50$)	BMVC'19	42.5	44.8	42.6	44.2	48.5	57.1	52.6	41.4	56.5	64.5	47.4	43.0	48.1	33.0	35.1	46.6
Yeh et al. (2019)	NIPS'19	44.8	46.1	43.3	46.4	49.0	55.2	44.6	44.0	58.3	62.7	47.1	43.9	48.6	32.7	33.3	46.7
Liu et al. (2020b) ($k = 243$)*	CVPR'20	41.8	44.8	41.1	44.9	47.4	54.1	43.4	42.2	56.2	63.6	45.3	43.5	45.3	31.3	32.2	45.1
UGCN (Wang et al., 2020) ($k = 96$)	ECCV'20	41.3	43.9	44.0	42.2	48.0	57.1	42.2	43.2	57.3	61.3	47.0	43.5	47.0	32.6	31.8	45.6
Chen et al. (2021) ($k = 81$)*	TCSVT'21	42.1	43.8	41.0	43.8	46.1	53.5	42.4	43.1	53.9	60.5	45.7	42.1	46.2	32.2	33.8	44.6
Zheng et al. (2021) ($k = 81$)*	ICCV'21	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
Li et al. (2022a) ($k = 351$)*	TMM'22	39.9	43.4	40.0	40.9	46.4	50.6	42.1	39.8	55.8	61.6	44.9	43.3	44.9	29.9	30.3	43.6
Xue et al. (2022) ($k = 243$)*	TIP'22	39.9	**42.7**	40.3	42.3	45.0	52.8	**40.4**	**39.3**	56.9	61.2	44.1	41.3	42.8	**28.4**	**29.3**	43.1
Li et al. (2022b) ($k = 351$)*	CVPR'22	39.2	43.1	40.1	40.9	44.9	51.2	40.6	41.3	53.5	60.3	43.7	**41.1**	43.8	29.8	30.6	43.0
Shan et al. (2022) ($k = 243$)*	ECCV'22	38.9	**42.7**	40.4	41.1	45.6	49.7	40.9	39.9	55.5	**59.4**	44.9	42.2	42.7	29.4	29.4	42.8
Zhao et al. (2023) ($k = 243$)*	CVPR'23	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	45.2
Chen et al. (2023) ($k = 96$)*	IJCAI'23	**38.1**	43.1	39.3	**39.4**	44.3	**49.1**	41.3	40.8	**53.1**	62.1	**43.3**	41.8	43.1	31.0	29.7	**42.6**
Ours ($k = 243$)*		39.0	43.4	**38.9**	41.0	**44.0**	50.7	40.8	40.6	54.1	60.3	43.7	41.4	**42.6**	28.9	**29.3**	**42.6**

* indicates that corresponding approaches take 2D poses detected by the cascaded pyramid network (CPN) as input. k represents the input frame number.

the fast branch, 9k frames supervision is enforced. Loss function can be calculated as follows:

$$L_{full} = \sum_{n=1}^{k}\sum_{i=1}^{M} \|P_i^{sn} - \overline{P}_i^{sn}\|_2 + \sum_{n=1}^{9k}\sum_{i=1}^{M} \|P_i^{fn} - \overline{P}_i^{fn}\|_2, \tag{16}$$

where P^s, P^f represent the ground-truth 3D joints and $\overline{P}^s, \overline{P}^f$ represent the predicted 3D joints. In addition, we propose a regulation item to ensure the prediction consistency between these two branches. The consistency regulation item firstly downsamples the 3D poses of fast branch and upsamples the 3D poses of slow branch. The formulation can be denoted as follows:

$$\tilde{P}^s = Down(\overline{P}^f), \tilde{P}^f = Up(\overline{P}^s). \tag{17}$$

The newly formed $\tilde{P}^s, \tilde{P}^f$ represent the supervision from the counterpart branch respectively. The consistency loss is calculated as:

$$L_r = \sum_{n=1}^{k}\sum_{i=1}^{M} \|\tilde{P}_i^{sn} - \overline{P}_i^{sn}\|_2 + \sum_{n=1}^{9k}\sum_{i=1}^{M} \|\tilde{P}_i^{fn} - \overline{P}_i^{fn}\|_2. \tag{18}$$

For the blending stage, only pose of central frame is predicted. In addition, each joint is supervised individually. Feature $\overline{Y} \in R^{M\times C}$ is used for joints prediction and a linear layer is enforced on $\overline{Y}$:

$$\overline{P} = Linear(\overline{Y}), \tag{19}$$

where predictions of each human joint is acquired. The loss function is calculated as follows:

$$L_{single} = \sum_{i=1}^{M} \|P_i - \overline{P}_i\|_2. \tag{20}$$

Overall loss function is calculated as:

$$L = L_{full} + L_{single} + \lambda L_r, \tag{21}$$

where λ is the balance weight for L_r. For inference, only pose $\overline{P}$ is used for evaluation.

4. Experiments

4.1. Datasets and evaluation metrics

We evaluate our model on three different benchmarks, i.e., Human3.6M (Ionescu et al., 2013), MPI-INF-3DHP (Mehta et al., 2017a) and HumanEva-I (Sigal et al., 2010).

Human3.6M. Human3.6M is one of the most representative benchmarks for 3D human pose estimation. The videos are captured in the indoor scene by 50 Hz cameras, where 3.6 million frames are saved. There are 11 actors participating in the shooting and 15 actions (e.g., greeting and sitting) are performed. We train the model on five human objects (S1, S5, S6, S7, S8) and test on two human objects (S9 and S11). For evaluation, two protocols are adopted, i.e., Protocol 1 and Protocol 2. Protocol 1 calculates the mean per joint position error (MPJPE). Protocol 2 calculates the Procrustes mean per joint position error (P-MPJPE). P-MPJPE firstly aligns the ground-truth pose and predicted pose via scale, translation and rotation. The position error is calculated afterwards.

MPI-INF-3DHP. MPI-INF-3DHP is another challenging benchmark for 3D human pose estimation. The dataset contains both the indoor and outdoor scenes, where 1.3 million images are provided. There are 8 actors participating in the shooting and 8 activities are included in the training set. The test set contains 7 activities. We only take 8 views from the training set for training and validate on the test set. We measure the MPJPE, percentage of correct keypoints with the threshold of 150 mm (PCK), and area under curve (AUC) for the evaluation of MPI-INF-3DHP dataset.

HumanEva-I. HumanEva-I is a smaller benchmark where only three human objects (S1, S2, S3) are involved. We report the results on three actions (Walk, Jog, Box) and only a single model is used for the evaluation. Following the common schedule, we adopt the P-MPJPE as the evaluation metric for this dataset.

4.2. Implementation details

Layer number of the transformer blocks is set the same as Zheng et al. (2021). For the training recipe, we employ the Adam optimizer with weight decay of 0.1. Total epochs of the training procedure are set to 80. We adopt cosine learning scheme where the initial learning rate is set to 1e-4 and the weight decay is 0.98. For Human 3.6M benchmark, we conduct experiments on both the detected and ground-truth 2D pose sequences. For MPI-INF-3DHP, ground-truth 2D pose sequences serve as the input of the framework. For HumanEva-I, 2D poses detected by MaskRCNN serve as the input source. We set hyperparameter λ in Eq. (21) as 0.1.

4.3. Quantitative evaluation

Human3.6M. Quantitative results on Human3.6M can be found in Tables 1 and 2. Inputs of 2D human joint results obtained from pretrained cascaded pyramid network are adopted. For Protocol 1, we can find that our approach achieves 42.6 mm MPJPE which is lower than all the previous approaches. Compared with MHFormer (Li et al., 2022b), our approach achieves lower MPJPE with less input frames (243 vs 351). We report the results under Protocol 2 in Table 2. We can observe that P-MPJPE of our approach is reduced to 34.2 mm. Promising results are achieved over different human actions.

When we take ground-truth 2D human poses as input, our approach can achieve the state-of-the-art results compared with the previous approaches. From Table 3 we can find that our approach reduces the MPJPE to 27.6 mm which is lower than all the previous approaches. Likewise, promising results are achieved over different human actions.

Table 2
Quantitative results of 3D HPE approaches and 'Ours' denotes the SlowFastFormer proposed in the paper. MPJPE under Protocol 2 (P-MPJPE) serves as the evaluation metric.

Protocol 2		Dir.	Disc.	Eat.	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Average
Pavlakos et al. (2018)	CVPR'18	34.7	39.8	41.8	38.6	42.5	47.5	38.0	36.6	50.7	56.8	42.6	39.6	43.9	32.1	36.5	41.8
Fang et al. (2018)	AAAI'18	38.2	41.7	43.7	44.9	48.5	55.3	40.2	38.2	54.5	64.4	47.2	44.3	47.3	36.7	41.7	45.7
Liu et al. (2020a)	ECCV'20	35.9	40.0	38.0	41.5	42.5	51.4	37.8	36.0	48.6	56.6	41.8	38.3	42.7	31.7	36.2	41.2
Zou and Tang (2021)	ICCV'21	35.7	38.6	36.3	40.5	39.2	44.5	37.0	35.4	46.4	51.2	40.5	35.6	41.7	30.7	33.9	39.1
Hossain and Little (2018)	ECCV'18	35.7	39.3	44.6	43.0	47.2	54.0	38.3	37.5	51.6	61.3	46.5	41.4	47.3	34.2	39.4	44.1
Cai et al. (2019) ($k = 7$)	ICCV'19	35.7	37.8	36.9	40.7	39.6	45.2	37.4	34.5	46.9	50.1	40.5	36.1	41.0	29.6	32.3	39.0
Lin and Lee (2019) ($k = 50$)	BMVC'19	32.5	35.3	34.3	36.2	37.8	43.0	33.0	32.2	45.7	51.8	38.4	32.8	37.5	25.8	28.9	36.8
Pavllo et al. (2019) ($k = 243$)*	CVPR'19	34.1	36.1	34.4	37.2	36.4	42.2	34.4	33.6	45.0	52.5	37.4	33.8	37.8	25.6	27.3	36.5
Liu et al. (2020b) ($k = 243$)*	CVPR'20	32.3	35.2	33.3	35.8	35.9	41.5	33.2	32.7	44.6	50.9	37.0	32.4	37.0	25.2	27.2	35.6
UGCN (Wang et al., 2020) ($k = 96$)	ECCV'20	32.9	35.2	35.6	34.4	36.4	42.7	**31.2**	32.5	45.6	50.2	37.3	32.8	36.3	26.0	23.9	35.5
Chen et al. (2021) ($k = 81$)*	TCSVT'21	33.1	35.3	33.4	35.9	36.1	41.7	32.8	33.3	**42.6**	49.4	37.0	32.7	36.5	25.5	27.9	35.6
Zheng et al. (2021) ($k = 81$)*	ICCV'21	32.5	**34.8**	32.6	34.6	**35.3**	39.5	32.1	32.0	42.8	**48.5**	**34.8**	**32.4**	35.3	24.5	26.0	34.6
Li et al. (2022a) ($k = 351$)*	TMM'22	32.7	35.5	32.5	35.4	35.9	41.6	33.0	31.9	45.1	50.1	36.3	33.5	35.1	23.9	25.0	35.2
Zhao et al. (2023) ($k = 243$)*	CVPR'23	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	35.6
Ours ($k = 243$)*		**31.6**	35.6	**31.7**	**33.4**	**35.3**	**38.7**	32.2	**31.7**	44.6	49.5	36.0	32.6	**33.7**	**23.3**	**24.7**	**34.2**

$*$ indicates that corresponding approaches take 2D poses detected by the cascaded pyramid network (CPN) as input. k represents the input frame number.

Table 3
Quantitative results of 3D HPE approaches and 'Ours' denotes the SlowFastFormer proposed in the paper. The approaches take the ground-truth 2D poses as input. MPJPE under Protocol 1 (MPJPE) serves as the evaluation metric.

Protocol 1		Dir.	Disc.	Eat.	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Average
P-LSTM (Lee et al., 2018) ($k = 243$)*	ECCV'18	32.1	36.6	34.3	37.8	44.5	49.9	40.9	36.2	44.1	45.6	35.3	35.9	30.3	37.6	35.5	38.4
Pavllo et al. (2019) ($k = 243$)*	CVPR'19	35.2	40.2	32.7	35.7	38.2	45.5	40.6	36.1	48.8	47.3	37.8	39.7	38.7	27.8	29.5	37.8
Liu et al. (2020b) ($k = 243$)*	CVPR'20	34.5	37.1	33.6	34.2	32.9	37.1	39.6	35.8	40.7	41.4	33.0	33.8	33.0	26.6	26.9	34.7
SRNet (Zeng et al., 2020) ($k = 96$)	ECCV'20	34.8	32.1	28.5	30.7	31.4	36.9	35.6	30.5	38.9	40.5	32.5	31.0	29.9	22.5	24.5	32.0
Chen et al. (2021) ($k = 81$)*	TCSVT'21	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	32.3
PoseAug (Gong et al., 2021) ($k = 81$)*	CVPR'21	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	38.2
Zheng et al. (2021) ($k = 81$)*	ICCV'21	30.0	33.6	29.9	31.0	30.2	33.3	34.8	31.4	37.8	38.6	31.7	31.5	29.0	23.3	23.1	31.3
Li et al. (2022b) ($k = 351$)*	CVPR'22	27.7	32.1	29.1	28.9	30.0	33.9	33.0	31.2	37.0	39.3	30.0	31.0	29.4	**22.2**	23.0	30.5
Ours ($k = 243$)*		**26.7**	**29.2**	**27.6**	**26.6**	**27.8**	**29.3**	**30.3**	**27.0**	**31.8**	**33.9**	**27.5**	**26.8**	**26.2**	**21.7**	**21.7**	**27.6**

k represents the input frame number.

Table 4
Quantitative results on MPI-INF-3DHP where PCK, AUC and MPJPE are reported. The best scores are marked in bold.

		PCK ↑	AUC ↑	MPJPE ↓
Mehta et al. (2017a)	3DV'17	75.7	39.3	117.6
Mehta et al. (2017b)	ACM ToG'17	76.6	40.4	124.7
Pavllo et al. (2019)	CVPR'19	86.0	51.9	84.0
Lin and Lee (2019)	BMVC'19	83.6	51.4	79.8
Li et al. (2020)	CVPR'20	81.2	46.1	99.7
Chen et al. (2021)	TCSVT'21	87.9	54.0	78.8
Zheng et al. (2021)	ICCV'21	88.6	56.4	77.1
Xue et al. (2022)	TIP' 22	90.3	57.8	69.4
Li et al. (2022b)	CVPR'22	93.8	63.3	58.0
Ours		**98.2**	**75.7**	**36.4**

Table 5
Quantitative results on HumanEva-I benchmark where Protocol 2 is employed for evaluation. The best scores are marked in bold.

	Walk			Jog			Box			Avg
	S1	S2	S3	S1	S2	S3	S1	S2	S3	
Pavlakos et al. (2017)	22.3	19.5	29.7	28.9	21.9	23.8	–	–	–	–
Martinez et al. (2017)	19.7	17.4	46.8	26.9	18.2	18.6	–	–	–	–
Pavlakos et al. (2018)	18.8	12.7	**29.2**	23.5	15.4	14.5	–	–	–	–
Lee et al. (2018)	18.6	19.9	30.5	25.7	16.8	17.7	42.8	48.1	53.4	30.3
Pavllo et al. (2019)	**13.9**	10.2	46.6	**20.9**	13.1	13.8	23.8	33.7	32.0	23.1
Zheng et al. (2021)	14.4	10.2	46.6	22.7	13.4	13.4	–	–	–	–
Ours (SA)	13.9	**9.5**	45.9	19.9	**12.9**	**12.6**	**22.9**	**33.2**	**29.1**	**22.2**

MPI-INF-3DHP. Quantitative results on MPI-INF-3DHP can be found in Table 4. We take the ground-truth 2D poses as input and achieve 98.2% PCK score which is a new state-of-the-art result. For the AUC and MPJPE metrics, we can still achieve state-of-the-art results, which verifies the effectiveness and generalization of the proposed approach in outdoor scenes.

HumanEva-I. Quantitative results on HumanEva-I can be found in Table 5. On this small dataset, our approach can achieve promising results under Protocol 2. We can observe that our approach can achieve lower P-MPJPE compared with previous approaches. Compared with PoseFormer (Zheng et al., 2021), we achieve lower error on *Walk* and *Jog*.

4.4. Ablation study

In this section, we will validate the effectiveness of each component proposed in the paper and the results are reported in Table 6. We set input frame number of slow and fast branches as 9, 81 separately. Firstly, we investigate the performance of slow or fast branch individually. From Table 6, we can find that slow branch only achieves 48.3 mm under MPJPE. In parallel, fast branch achieves 47.9 mm under MPJPE. When we build the two stream framework without any additional module, the prime slow-fast framework reduces the MPJPE to 47.5 mm. We can conclude that the slow-fast architecture brings non-trivial improvement over network of single branch. Different kinds

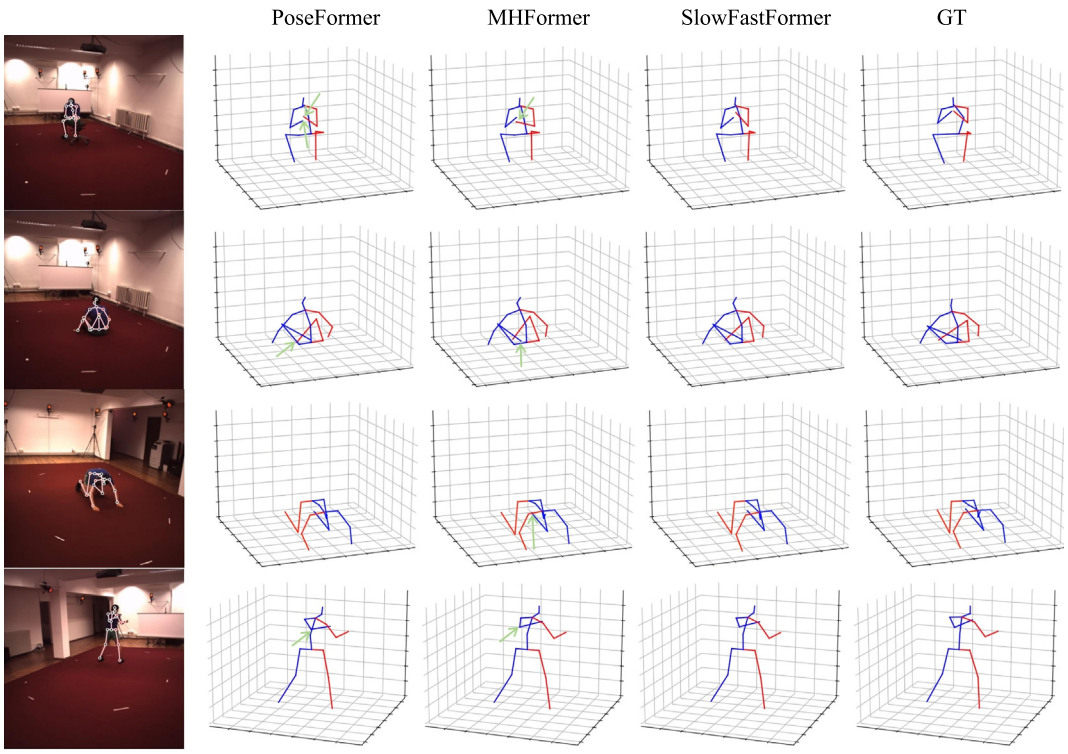

Fig. 3. Visualizations of estimated poses on subject S11 of Human3.6M dataset where different kinds of actions are included.

Table 6
Comparison on different components of our model. Slow means the branch where input frame number is 9 while Fast means the branch where frame number is 81. SF means the network with parallel encoding only. SF-LA means the Slow-Fast network equipped with lateral connection. SF-LA-B incorporated the blending module based on SF-LA. H1 means hierarchical supervision scheme without the consistency regulation item and H2 means hierarchical supervision scheme equipped with consistency item.

Approach	MPJPE (mm)
Slow	48.3
Fast	47.9
SF	47.5
SF-LA	47.3
SF-LA-B	46.9
SF-LA-B-H1	46.7
SF-LA-B-H2	46.1

Table 7
Comparison on lateral connection schemes of our model. SF-LA means the prime lateral connection scheme. SF-LAS means the lateral connection equipped with hyperparameter. SF-LAC means the lateral connection where convolution is involved.

Approach	MPJPE (mm)
SF-LA	47.3
SF-LAS	47.9
SF-LAC	48.3

Table 8
Comparison on different regulation schemes. SF-LA-B-H3 and SF-LA-B-H4 means the other versions of regulation item.

Approach	MPJPE (mm)
SF-LA-B-H2	46.1
SF-LA-B-H3	46.9
SF-LA-B-H4	46.6

of contexts make great significance to the pose estimation in videos. Lateral connection can bridge the gap between the parallel branches and information exchange benefits the feature enhancement of separate branches. From the table, we can observe that lateral connection reduces the MPJPE to 47.3 mm. Based on the first stage, if blending stage is incorporated, the error drops to 46.9 mm. Hierarchical supervision can refine the predictions progressively and the two-level supervision makes the network speed-aware and joint-aware. From the table we can find that the supervision scheme without the regulation item reduces the error to 46.7 mm. When adding the consistency loss, the MPJPE is reduced to 46.1 mm which proves the effectiveness of prediction consistency between fast and slow branches.

Impact of different lateral connection schemes. Table 7 demonstrates the effect of different lateral connection schemes. SF-LA is our prime version which achieves the best performance. When combining two branches with hyperparameters, the MPJPE increases to 47.9 mm. It is hard for the network to learn appropriate hyperparameters weighting different paths. When adding convolution operation along the lateral path, the MPJPE is 48.3 mm and the prime lateral connection scheme is the best choice.

Impact of different supervision schemes. Table 8 demonstrates the performance of different supervision schemes. SF-LA-B-H2 denotes the supervision scheme mentioned in the maintext. In this version, regulation items $\tilde{P}^s, \tilde{P}^f$ are detached from the main network. For SF-LA-B-H3, regulation items $\tilde{P}^s, \tilde{P}^f$ are not detached. We can find that SF-LA-B-H2 achieves better performance. In SF-LA-B-H4, only f frames is supervised for the fast branch. We can observe that the MPJPE increases to 46.6 mm and the full supervision over fast path is necessary.

Efficiency analysis. Table 9 analyses the computational complexity of the whole framework. We can observe that our approach achieves the lowest projection error with the smallest computational complexity. The FLOPs is only 740M when input frame number is 243, while FLOPs of P-STMO (Shan et al., 2022) reaches 1737M. For MHFormer (Li et al.,

Table 9
Complexity comparisons of different approaches.

Approach	k	FLOPs (M)	MPJPE (mm)
PoseFormer (Zheng et al., 2021) ICCV'21	81	1624	44.3
MHFormer (Li et al., 2022b) CVPR'22	27	1031	45.9
StridedFormer (Li et al., 2022a) TMM'22	351	2142	43.7
P-STMO (Shan et al., 2022) ECCV'22	243	1737	42.8
SlowFastFormer	243	**740**	**42.6**

'k' represents the frame number of input.

2022b), we can find that FLOPs reaches 1031M when frame number is 27. However, the MPJPE of MHFormer reaches 45.9 mm which is much higher than ours. From the analysis above we can conclude that our approach with lower computational complexity is more efficient.

4.5. Qualitative results

Qualitative results can be found in Fig. 3. We mainly visualize the poses of actions including *Sitting Down, Sitting,* etc. From the figure we can find that the approach can achieve promising results over these complex poses compared with PoseFormer (Zheng et al., 2021) and MHFormer (Li et al., 2022b).

5. Conclusion

In this paper, we firstly propose a slow-fast framework for 3D human pose estimation based on transformer and we dub the network as SlowFastFormer. Firstly, a parallel encoding module is proposed to encode different kinds of context from slow and fast branches. Secondly, to promote the message passing among different branches, a blending module is tailored where message exchange is performed. These two modules enhance the features in a progressive manner. Lastly, the hierarchical supervision scheme refines the predictions progressively and network is speed-aware through the intermediate supervision. Our approach achieves leading performance on several benchmarks which proves the effectiveness of SlowFastFormer.

CRediT authorship contribution statement

Lu Zhou: Conceptualization, Investigation, Methodology, Writing – original draft, Validation, Visualization. **Yingying Chen:** Funding acquisition, Resources, Supervision, Writing – review & editing. **Jinqiao Wang:** Funding acquisition, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China (No. 2022ZD0160601). This work was supported by National Natural Science Foundation of China under Grants 62206283, 62276260, 62076235 and 62176254.

References

Ahn, J., Jang, Y., Chung, J.S., 2023. SlowFast network for continuous sign language recognition. arXiv preprint arXiv:2309.12304.
Anon, 2023a. Collaborative three-stream transformers for video captioning. Comput. Vis. Image Underst. 235, 103799.
Anon, 2023b. Global–local contrastive multiview representation learning for skeleton-based action recognition. Comput. Vis. Image Underst. 229, 103655.
Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T.-J., Yuan, J., Thalmann, N.M., 2019. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In: Proc. IEEE Int. Conf. Comput. Vis. pp. 2272–2281.
Cai, J., Liu, H., Ding, R., Li, W., Wu, J., Ban, M., 2023. HTNet: Human topology aware network for 3d human pose estimation. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. IEEE, pp. 1–5.
Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. In: Proc. Eur. Conf. Comput. Vis. pp. 213–229.
Chen, T., Fang, C., Shen, X., Zhu, Y., Chen, Z., Luo, J., 2021. Anatomy-aware 3d human pose estimation with bone-based pose decomposition. IEEE Trans. Circuits Syst. Video Technol. 32 (1), 198–209.
Chen, H., He, J.-Y., Xiang, W., Liu, W., Cheng, Z.-Q., Liu, H., Luo, B., Geng, Y., Xie, X., 2023. HDFormer: High-order directed transformer for 3D human pose estimation. In: Int. Joint Conf. Artif. Intell.
Chu, X., Tian, Z., Zhang, B., Wang, X., Wei, X., Xia, H., Shen, C., 2021. Conditional positional encodings for vision transformers. arXiv preprint arXiv:2102.10882.
Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929.
Fang, H.-S., Xu, Y., Wang, W., Liu, X., Zhu, S.-C., 2018. Learning pose grammar to encode human body configuration for 3d pose estimation. In: Proc. IEEE Int. Conf. Comput. Vis., Vol. 32, No. 1.
Fayyaz, M., Koohpayegani, S.A., Jafari, F.R., Sengupta, S., Joze, H.R.V., Sommerlade, E., Pirsiavash, H., Gall, J., 2022. Adaptive token sampling for efficient vision transformers. In: Proc. Eur. Conf. Comput. Vis. Springer, pp. 396–414.
Feichtenhofer, C., Fan, H., Malik, J., He, K., 2019. Slowfast networks for video recognition. In: Proc. IEEE Int. Conf. Comput. Vis. pp. 6202–6211.
Foo, L.G., Li, T., Rahmani, H., Ke, Q., Liu, J., 2023. Unified pose sequence modeling. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 13019–13030.
Garcia-Salguero, M., Gonzalez-Jimenez, J., Moreno, F.-A., 2019. Human 3D pose estimation with a tilting camera for social mobile robot interaction. Sensors 19 (22), 4943.
Gedamu, K., Ji, Y., Gao, L., Yang, Y., Shen, H.T., 2023. Relation-mining self-attention network for skeleton-based human action recognition. Pattern Recognit. 139, 109455.
Gong, K., Zhang, J., Feng, J., 2021. Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 8575–8584.
Gui, L.-Y., Zhang, K., Wang, Y.-X., Liang, X., Moura, J.M., Veloso, M., 2018. Teaching robots to predict human motion. In: Proc. Int. Conf. Intell. Robots Syst. pp. 562–567.
Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y., 2021. Transformer in transformer. Proc. Adv. Neural Inform. Process. Syst. 34, 15908–15919.
Han, C., Yu, X., Gao, C., Sang, N., Yang, Y., 2022. Single image based 3D human pose estimation via uncertainty learning. Pattern Recognit. 132, 108934.
Hossain, M.R.I., Little, J.J., 2018. Exploiting temporal information for 3d human pose estimation. In: Proc. Eur. Conf. Comput. Vis. pp. 68–84.
Hu, W., Zhang, C., Zhan, F., Zhang, L., Wong, T.-T., 2021. Conditional directed graph convolution for 3d human pose estimation. In: Proc. ACM Int. Conf. Multimedia. pp. 602–611.
Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C., 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. 36 (7), 1325–1339.
Jiang, C., Huang, K., Zhang, S., Wang, X., Xiao, J., Goulermas, Y., 2023. Aggregated pyramid gating network for human pose estimation without pre-training. Pattern Recognit. 138, 109429.
Kang, Y., Liu, Y., Yao, A., Wang, S., Wu, E., 2023. 3D human pose lifting with grid convolution. arXiv preprint arXiv:2302.08760.

Kazakos, E., Nagrani, A., Zisserman, A., Damen, D., 2021. Slow-fast auditory streams for audio recognition. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. IEEE, pp. 855–859.

Lee, K., Lee, I., Lee, S., 2018. Propagating lstm: 3d pose estimation based on joint interdependency. In: Proc. Eur. Conf. Comput. Vis. pp. 119–135.

Li, S., Ke, L., Pratama, K., Tai, Y.-W., Tang, C.-K., Cheng, K.-T., 2020. Cascaded deep monocular 3D human pose estimation with evolutionary training data. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 6173–6183.

Li, W., Liu, H., Ding, R., Liu, M., Wang, P., Yang, W., 2022a. Exploiting temporal contexts with strided transformer for 3d human pose estimation. IEEE Trans. Multimedia.

Li, W., Liu, H., Tang, H., Wang, P., Van Gool, L., 2022b. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 13147–13156.

Li, Y., Mao, H., Girshick, R., He, K., 2022c. Exploring plain vision transformer backbones for object detection. arXiv preprint arXiv:2203.16527.

Li, K., Wang, S., Zhang, X., Xu, Y., Xu, W., Tu, Z., 2021. Pose recognition with cascade transformers. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 1944–1953.

Lin, J., Lee, G.H., 2019. Trajectory space factorization for deep video-based 3d human pose estimation. arXiv:1908.08289.

Liu, J., Ding, H., Shahroudy, A., Duan, L.-Y., Jiang, X., Wang, G., Kot, A.C., 2019. Feature boosting network for 3D pose estimation. IEEE Trans. Pattern Anal. Mach. Intell. 42 (2), 494–501.

Liu, K., Ding, R., Zou, Z., Wang, L., Tang, W., 2020a. A comprehensive study of weight sharing in graph networks for 3d human pose estimation. In: Proc. Eur. Conf. Comput. Vis. pp. 318–334.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proc. IEEE Int. Conf. Comput. Vis. pp. 10012–10022.

Liu, R., Shen, J., Wang, H., Chen, C., Cheung, S.-c., Asari, V., 2020b. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 5064–5073.

Martinez, J., Hossain, R., Romero, J., Little, J.J., 2017. A simple yet effective baseline for 3d human pose estimation. In: Proc. IEEE Int. Conf. Comput. Vis. pp. 2640–2649.

Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C., 2017a. Monocular 3d human pose estimation in the wild using improved cnn supervision. In: Proc. Int. Conf. 3D Vis. pp. 506–516.

Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.-P., Xu, W., Casas, D., Theobalt, C., 2017b. Vnect: Real-time 3d human pose estimation with a single rgb camera. ACM Trans. Graph. 36 (4), 1–14.

Pan, J., Bulat, A., Tan, F., Zhu, X., Dudziak, L., Li, H., Tzimiropoulos, G., Martinez, B., 2022. Edgevits: Competing light-weight cnns on mobile devices with vision transformers. In: Proc. Eur. Conf. Comput. Vis. Springer, pp. 294–311.

Pavlakos, G., Zhou, X., Daniilidis, K., 2018. Ordinal depth supervision for 3d human pose estimation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 7307–7316.

Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K., 2017. Coarse-to-fine volumetric prediction for single-image 3D human pose. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 7025–7034.

Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M., 2019. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 7753–7762.

Peng, W., Hong, X., Zhao, G., 2021. Tripool: Graph triplet pooling for 3D skeleton-based action recognition. Pattern Recognit. 115, 107921.

Shan, W., Liu, Z., Zhang, X., Wang, S., Ma, S., Gao, W., 2022. P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In: Proc. Eur. Conf. Comput. Vis. Springer, pp. 461–478.

Sharma, S., Varigonda, P.T., Bindal, P., Sharma, A., Jain, A., 2019. Monocular 3d human pose estimation by generation and ordinal ranking. In: Proc. IEEE Int. Conf. Comput. Vis. pp. 2325–2334.

Sigal, L., Balan, A.O., Black, M.J., 2010. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. Int. J. Comput. Vis. 87 (1–2), 4.

Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep high-resolution representation learning for human pose estimation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 5693–5703.

Tang, Z., Qiu, Z., Hao, Y., Hong, R., Yao, T., 2023. 3D human pose estimation with spatio-temporal criss-cross attention. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 4790–4799.

Tian, L., Wang, P., Liang, G., Shen, C., 2021. An adversarial human pose estimation network injected with graph structure. Pattern Recognit. 115, 107863.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Proc. Adv. Neural Inf. Proces. Syst. pp. 5998–6008.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L., 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proc. IEEE Int. Conf. Comput. Vis. pp. 568–578.

Wang, C., Xu, H., Zhang, X., Wang, L., Zheng, Z., Liu, H., 2022. Convolutional embedding makes hierarchical vision transformer stronger. In: Proc. Eur. Conf. Comput. Vis. Springer, pp. 739–756.

Wang, J., Yan, S., Xiong, Y., Lin, D., 2020. Motion guided 3d pose estimation from videos. In: Proc. Eur. Conf. Comput. Vis. pp. 764–780.

Wei, G., Lan, C., Zeng, W., Chen, Z., 2019. View invariant 3D human pose estimation. IEEE Trans. Circuits Syst. Video Technol. 30 (12), 4601–4610.

Wei, D., Tian, Y., Wei, L., Zhong, H., Chen, S., Pu, S., Lu, H., 2022. Efficient dual attention slowfast networks for video action recognition. Comput. Vis. Image Underst. 222, 103484.

Wu, L., Zhang, C., Zou, Y., 2023. SpatioTemporal focus for skeleton-based action recognition. Pattern Recognit. 136, 109231.

Xu, T., Takano, W., 2021. Graph stacked hourglass networks for 3D human pose estimation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 16105–16114.

Xue, Y., Chen, J., Gu, X., Ma, H., Ma, H., 2022. Boosting monocular 3D human pose estimation with part aware attention. IEEE Trans. Image Process. 31, 4278–4291.

Yeh, R., Hu, Y.-T., Schwing, A., 2019. Chirality nets for human pose regression. Proc. Adv. Neural Inf. Proces. Syst. 32, 8163–8173.

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.-H., Tay, F.E., Feng, J., Yan, S., 2021a. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: Proc. IEEE Int. Conf. Comput. Vis. pp. 558–567.

Yuan, Y., Fu, R., Huang, L., Lin, W., Zhang, C., Chen, X., Wang, J., 2021b. Hrformer: High-resolution vision transformer for dense predict. Proc. Adv. Neural Inform. Process. Syst. 34, 7281–7293.

Yuan, Z., Xue, C., Chen, Y., Wu, Q., Sun, G., 2022. Ptq4vit: Post-training quantization for vision transformers with twin uniform quantization. In: Proc. Eur. Conf. Comput. Vis. Springer, pp. 191–207.

Zeng, A., Sun, X., Huang, F., Liu, M., Xu, Q., Lin, S., 2020. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In: Proc. Eur. Conf. Comput. Vis. pp. 507–523.

Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N., 2019. Semantic graph convolutional networks for 3d human pose regression. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 3425–3435.

Zhao, Q., Zheng, C., Liu, M., Wang, P., Chen, C., 2023. PoseFormerV2: Exploring frequency domain for efficient and robust 3D human pose estimation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 8877–8886.

Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z., 2021. 3D human pose estimation with spatial and temporal transformers. arXiv:2103.10455.

Zong, Z., Li, K., Song, G., Wang, Y., Qiao, Y., Leng, B., Liu, Y., 2022. Self-slimmed vision transformer. In: Proc. Eur. Conf. Comput. Vis. Springer, pp. 432–448.

Zou, Z., Tang, W., 2021. Modulated graph convolutional network for 3d human pose estimation. In: Proc. IEEE Int. Conf. Comput. Vis. pp. 11477–11487.