Multimedia Tools and Applications (2024) 83:73533–73552
https://doi.org/10.1007/s11042-023-17955-6

1232: HUMAN-CENTRIC MULTIMEDIA ANALYSIS

3D Human pose estimation from video via multi-scale multi-level spatial temporal features

Liling Fan[1,2] · Kunliang Jiang[1,2] · Weixue Zhou[1,2] · Zhenguo Gao[1,2] · Yanmin Luo[1]

Received: 27 March 2023 / Revised: 6 October 2023 / Accepted: 18 December 2023 /
Published online: 22 January 2024

Abstract

In this paper, we present an innovative framework for 2D-to-3D human pose estimation from video, harnessing the power of multi-scale multi-level spatial-temporal features. Our framework comprises three integral branch networks: A temporal feature core network, dedicated to extracting temporal coherence among frames, enabling a comprehensive understanding of dynamic human motion. A multi-scale feature branch network, equipped with multiple receptive fields of varying sizes, facilitating the extraction of multi-scale features, thus capturing fine-grained details across different scales. A multi-level feature branch network, tasked with extracting features from layers at various depths within the architecture, providing a nuanced understanding of pose-related information. Within our framework, these diverse features are seamlessly integrated to encapsulate intricate spatial and temporal relationships inherent to the human body. This integration effectively addresses challenges such as depth ambiguity and self-occlusions, culminating in substantially improved accuracy in pose estimation.Extensive experiments on Human3.6M and HumanEva-I show that our framework achieves competitive performance on 2D-to-3D human pose estimation in video. Code is available at: https://github.com/fll123/3Dhumanpose.

Keywords 2D-to-3D Human pose estimation · Spatial-wise separable residual convolution · Multi-scale multi-level spatial temporal feature

✉ Zhenguo Gao
gaohit@sina.com

Liling Fan
975212384@qq.com

Kunliang Jiang
jiang_kl@foxmail.com

Weixue Zhou
2997977@qq.com

Yanmin Luo
lym@hqu.edu.cn

[1] College of Computer Science and Technology, Huaqiao University, Xiamen 361021, Fujian, China

[2] Key Laboratory of Computer Vision and Machine Learning, Fujian Province University, Xiamen 361021, Fujian, China

1 Introduction

In recent years, 3D human pose estimation has garnered significant attention due to its promising applications in behavior recognition [1] and pose tracking [2]. Traditional approaches to 3D human pose estimation, such as [3] and marker systems [4], have historically relied on specialized hardware and meticulous setup procedures within controlled environments. These requirements have imposed significant hardware demands and labor-intensive processes, thereby hindering the widespread adoption of 3D human pose estimation solutions [5]. However, recent strides in deep learning have transformed this landscape, enabling efficient 3D human pose estimation from standard videos and images. This paradigm shift not only obviates the need for resource-intensive devices and elaborate setups but also propels the state-of-the-art performance of 3D human pose estimation [6, 7].

Deep learning-based methods for 3D human pose estimation can generally be classified into two approaches: an end-to-end approach and a two-stage approach. The end-to-end approach relies on deep neural networks to directly predict 3D human poses in an image [8–10]. Constructively, the two-stage approach first predicts 2D human joint positions using dedicated 2D pose estimation methods, and then elevates 2D joint positions to 3D positions through regression networks [11–14]. Recently, rapid advances in 2D human pose estimation make the two-step approach a promising approach, which makes 2D-to-3D mapping a hot topic. 2D-to-3D methods leverage 2D key points to achieve high accuracy.

Many early works on 3D human pose estimation [15, 16] rely on single frame image analysis. However, as a combination of multiple consecutive frames, a video contains more complex temporal information. As a result, a single image-based estimation method may lead to large estimation differences between adjacent frames. In fact, temporal incoherence and jitter are often obtained from single frame estimations.

For this reason, to produce more robust estimations, some scholars have attempted to estimate the human pose in a particular frame from a sequence of frames in a video. Recurrent neural network (RNN) is a widely used model that is powerful for modeling sequence data. Some work [12, 17] uses RNNs to construct regression models for extracting temporal information from sequence data. Many results have shown that 3D human pose estimation based on sequence data is more accurate than a single frame. However, RNNs are prone to gradient vanishing and explosion problems when dealing with long sequence data. Pavllo et al. [14] design a temporal convolutional network (TCN) for 3D human pose estimation from 2D joint motion trajectories. TCN can flexibly capture various-length sequences and can support causal convolutions to allow online estimation. TCN has yielded dramatic improvements in 2D-to-3D human pose estimation methods.

2D-to-3D human pose estimation is still an ill-posed problem for reasons including depth ambiguity and self-occlusions. Many works [11, 14, 18] adopt cascaded multi-layer network architectures, whereas not benefited from model depth. Moreover, the features extracted by this architecture are simple, which limits interpretability of the model.

In this work, focusing on the two-step approach to 3D human pose estimation from video, we propose a new framework for 2D-to-3D mapping. The framework contains three branch networks: a temporal feature core network for exploiting temporal coherence among frames, a multi-scale feature branch network for exploiting multi-scale features using multiple receptive fields of various sizes, and a multi-level feature branch network for exploiting multi-level features from layers at different levels. Within our framework, these diverse features are meticulously exploited to resolve challenges such as depth ambiguity and self-occlusions, leading to significantly more accurate 3D human pose estimations derived from 2D skeleton key points.

A cornerstone of our framework is the spatial-wise separable residual convolution module, which builds upon the depth-wise separable convolution module introduced in [19]. Notable improvements in our module include: (1) the strategic swapping of the order between depth-wise convolution and point-wise convolution, (2) the replacement of the ReLU activation function with the superior Mish function, and (3) the incorporation of a residual connection. Empirical results unequivocally demonstrate that our enhanced module delivers superior performance while requiring fewer parameters.

Extensive experiments conducted on two publicly available datasets affirm the efficacy of our proposed framework. It not only enhances the accuracy of 3D human pose estimation but also consistently achieves state-of-the-art performance in the field.

In summary, our main contributions to this paper are as follows:

(1) We propose three branch networks for extracting multi-scale features, multi-level features, and temporal coherence features for estimating human pose from videos.
(2) We propose a new framework for 2D-to-3D mapping, which achieves 3D human pose estimations from 2D skeleton key points by fusing multi-scale, multi-level spatial temporal features.
(3) Extensive experiments on Human3.6M and HumanEva-I show that the proposed method generally achieves competitive performance.

2 Related work

2.1 Efficient architecture

Convolutional neural networks have been widely used in many scenarios, such as image analysis [20], action recognition [21, 22], and human pose estimation [23, 24]. As the demands for more complex prediction tasks have grown, CNNs have evolved towards increasingly intricate and deeper architectures, presenting significant computational challenges. To address this issue, a wave of lightweight CNNs has recently emerged.

He et al. introduce ResNet [25], a network incorporating residual connections to mitigate the vanishing gradient problem and expedite network convergence. Sifre et al. propose depth-wise separable convolution [19], which decomposes computation into depth-wise and point-wise convolutions, simplifying standard convolution parameters, thereby reducing model complexity. MobileNetV2[26] introduce inverted residual blocks with fewer parameters and lower computational costs on the basis of depth-wise separable convolution[19]. In line with these advancements, we also focus on depth-wise separable convolution and present a novel convolution structure tailored for 3D human pose tasks.

2.2 3D Human pose estimation

In the realm of 3D human pose estimation using deep learning, two predominant approaches have emerged as the mainstay: the end-to-end approach and the two-stage approach. The end-to-end approach relies on a regression network to directly predict 3D human poses from input images [8, 9, 27]. Notably, Li et al. [27] introduce a multi-task convolutional neural network (CNN) that leverages shared network results for joint prediction and detection, achieving end-to-end 3D human pose estimation. Similarly, Nie et al. [8] propose a Structured Pose Representation (SPR) that effectively combines person instance and body joint position representations. Building upon SPR, they developed the SPM model, which directly predicts poses in an end-to-end fashion.

In contrast, the two-stage approach proceeds with a more structured pipeline, initially estimating 2D joint positions and subsequently elevating them to 3D positions through a regression network [11–14]. This method often attains higher average accuracy compared to the end-to-end approach, prompting its widespread adoption. For instance, Martinez et al. [11] introduce a straightforward yet highly effective model for 3D human pose estimation, achieving impressive accuracy through fully connected layers augmented with residual connections. Hossain et al. [12] employ sequence-to-sequence LSTM networks to extract temporal information and enforced temporal smoothness constraints during training. Recently, Zhao et al. [18] introduce semantic Graph Convolutional Networks (GCN) to enhance graph convolution performance by learning weights for implicit prior edges in a graph, further improving the performance of two-stage methods. Additionally, Pavllo et al. [14] harness a temporal convolutional network composed of dilated convolutions to capture temporal information from extended sequences, leading to enhanced performance in 3D human pose estimation.

3 Multi-scale multi-level spatial temporal feature network

In this section, we introduce our multi-scale multi-level spatial temporal feature network for 2D to 3D human pose estimation. The network generates a sequence of 3D coordinates rooted at the pelvis root joint from a sequence of 2D pose predictions from a video. The 2D to 3D pose mapping process is illustrated in Fig. 1.

3.1 Structure of our network

Figure 2 depicts the overall architecture of our network, which consists of three branch networks, as outlined by the three dashed rectangles. The branch network in the blue dashed rectangle is the core network for extracting temporal features. The branch network in the orange dashed rectangle is the multi-scale feature branch network for extracting multi-scale features. The branch network in the orange dashed rectangle is the multi-level feature branch network for extracting multi-level features.

The three branch networks are build on basic modules, including Expand_Conv module, SSDR_Conv module and SSSR_Conv module. The structures of the basic modules are shown in Fig. 3.

The multi-scale feature branch network, which performs multi-scale processing on the input and learns more feature representations, is built on expanded convolution

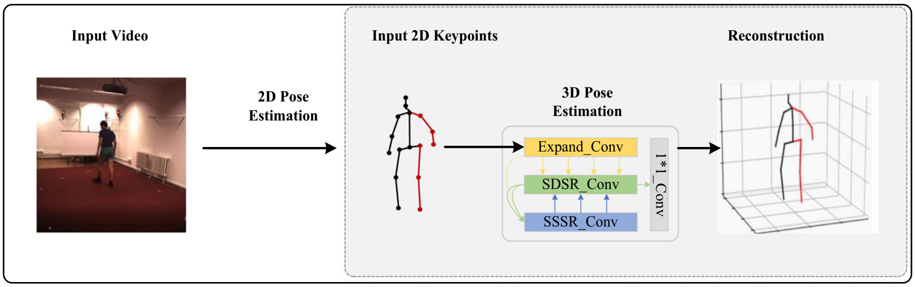

Fig. 1 Schematic of the 2D to 3D mapping. The input is a skeleton with 2D key points, and the output is a 3D pose reconstructed from the corresponding 2D key points

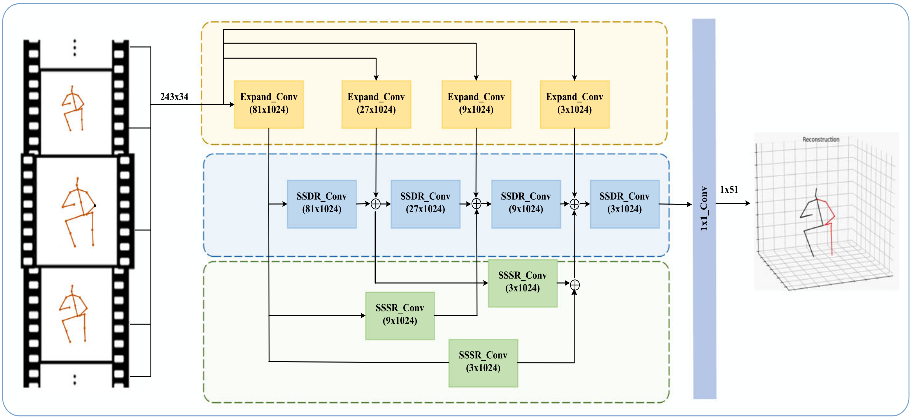

Fig. 2 An instantiation of our network architecture for 3D human pose estimation. The input consists of 2D key points for a receptive field of 243 frames with J = 17 joints. The tensor sizes as shown in parentheses, e.g., (81x1024) denote 81 frames and 1024 channels

(Expand_Conv) modules. An Expand_Conv module consists of a 1D convolution with kernel size **k** and stride factor **s**. The temporal feature core network contains spatial separable dilated residual convolution (SSDR_Conv) modules to capture long-range temporal relationships across frames, thereby significantly enhancing temporal consistency. A SSDR_Conv module is built on SSDR_Conv blocks with a dilation factor **d**. The multi-level feature branch network consists of spatial separable strided residual convolution (SSSR_Conv) modules, which refines the shallow features of the model to obtain higher-level features. A SSSR_Conv module is built on SSSR_Conv blocks with a stride factor **s**.

In our network architecture, each block is followed by a 1D convolution with kernel size 1. We adopt the temporal feature core network as the baseline model.

3.2 Spatial-wise separable residual convolution

Inspired by depth-wise separable convolution, we introduce spatial-wise separable residual convolution to capture temporal information of long sequences.

Depth-wise separable convolution involves both spatial dimension and depth dimension. As shown in Fig. 4, its operation contains two parts, depth-wise convolution and point-wise convolution. Depth-wise convolution is a separate convolution operation on each channel of the input image, which extracts features from each channel. Point-wise convolution is a spatial 1x1 convolution operation for fusing the feature map across channels. Splitting traditional convolution into depth-wise convolution and point-wise convolution reduces the number of convolution parameters.

However, depth-wise separated convolution does not perform as well as standard convolution, as features extracted from each channel may lose some valuable information, leading to a degradation in performance. To mitigate this negative effect, we make some modifications: we swap the order of depth-wise convolution and point-wise convolution, use Mish function to replace the ReLU activation function, and add a residual connection. The modifications do not bring additional parameters.

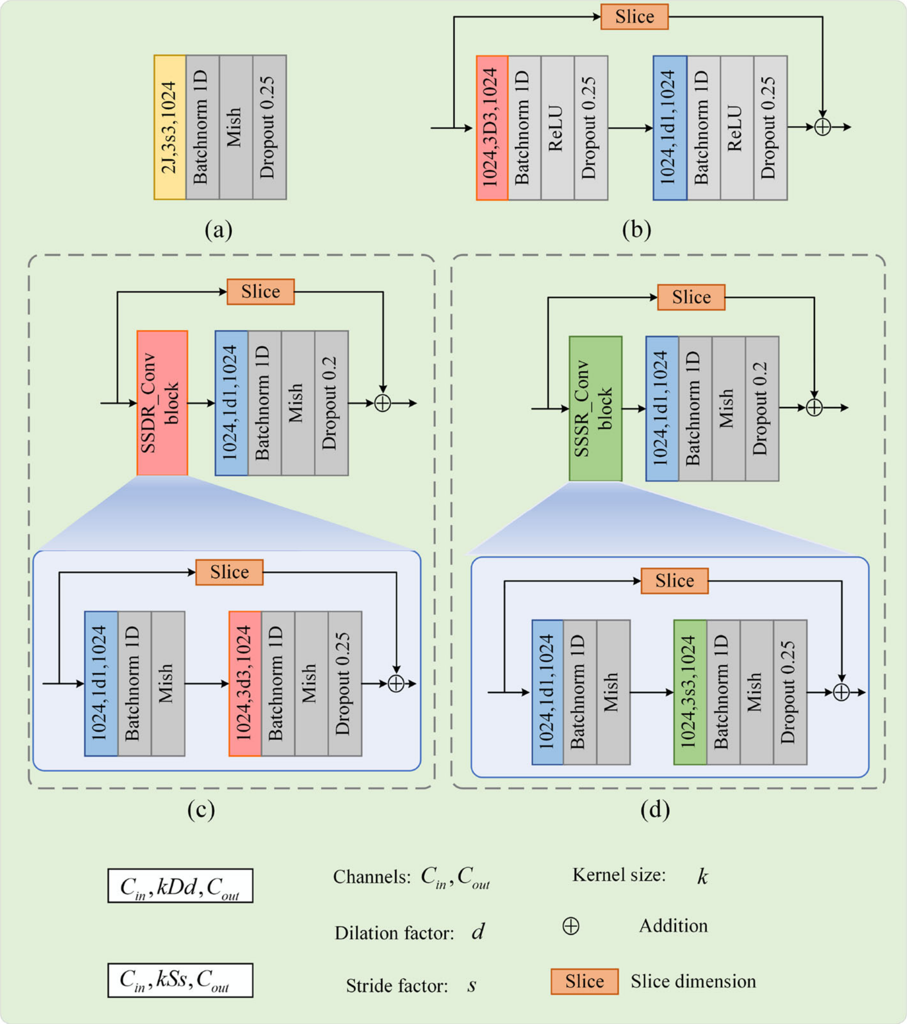

Fig. 3 The component modules. **(a)** The Expand_Conv module; **(b)** The basic block of TCN; **(c)** The spatial-wise separable residual convolution module parameterized with dilation; **(d)** The spatial-wise separable residual convolution module parameterized with stride

The main significance of activation functions in neural networks is to introduce nonlinearities that enable the network to learn complex patterns in the data. A commonly used activation function is the rectified linear activation function (ReLU) [28] defined as:

$$f(x) = max(0, x). \tag{1}$$

ReLU has good non-saturation properties, but it directly sets negative values to zero, resulting in a certain amount of feature loss. To compensate for it, we adopt the non-monotonic neural activation function (Mish), which is defined as:

$$f(x) = x\ tanh(\ln(1 + e^x)) \tag{2}$$

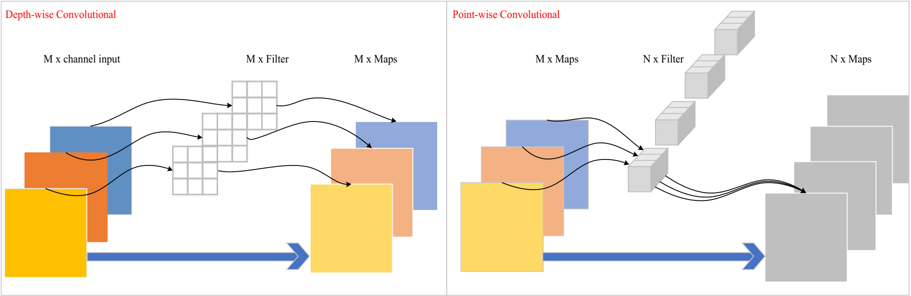

Fig. 4 The process of depth-wise separable convolution. This process includes depth-wise convolution and point-wise convolution

Mish and Relu are compared in Fig. 5. The Mish function has the advantages of being non-monotonic and smooth. A smooth activation function allows information to better percolate into the neural network, resulting in better accuracy and generalization. Therefore, to predict more accurate 3D human poses, we use Mish function in our modules.

For facilitating the comparison, the structures of the depth-wise separable convolution and our spatial-wise separable residual convolution are respectively shown in the sub-figures of Fig. 6. In our network, the SSDR_Conv blocks adopt our spatial-wise separable residual convolution parameterized with a dilation factor, and the SSSR_Conv blocks use our spatial-wise separable residual convolution parameterized with a stride factor.

3.3 Multi-scale and multi-level features

We use multi-scale and multi-level features to construct structures with powerful multi-scale feature extraction capabilities to enable highly accurate 2D to 3D human pose estimation.

Multi-scale Feature. Many previous works [29–31] have found that multi-scale features enable models to learn both local and global features. Using deep learning for 3D human pose estimation, multi-scale features are crucial to pose understanding.

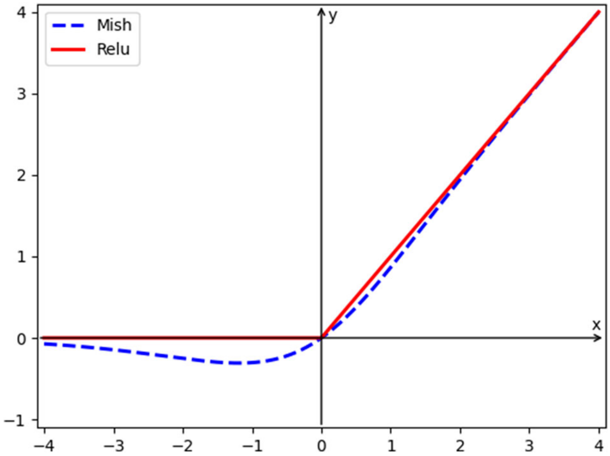

Fig. 5 Comparison of Mish and Relu activation functions

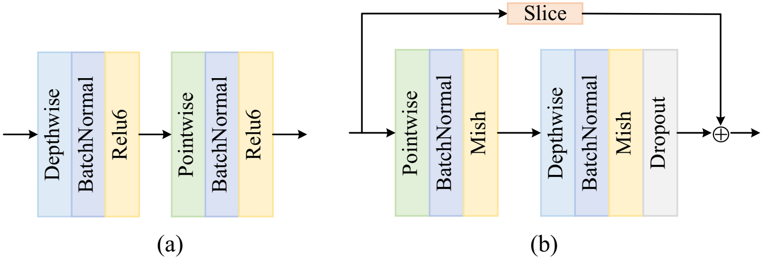

Fig. 6 Comparison of the structure of depth-wise separable convolution and our convolutional structure. **(a)** Structure of depth-wise separable convolution; **(b)** Structure of our spatial-wise separable residual convolution

Multi-level Feature. Newell et al. [32] found that features in the shallow layers of the network are equally valuable for research. Therefore, we extract multi-level features in shallow layers and believe that this can also bring valuable information for the final prediction.

As shown in Fig. 7, we use the Expand_Conv module to upgrade the channel of input features and process multi-scale receptive fields in the multi-scale feature branch network. Features are fused to each layer of the network to propagate richer features. In the multi-level feature branch network, we use the SSSR_Conv module to refine the output of the 0th SSDR_Conv layer and the 0th Expand_Conv layer to obtain higher-level features that accumulated into the subsequent layers of the model.

4 Experiments

4.1 DataSets and evaluation protocols

To validate the effectiveness of the proposed method, we conduct experiments on two motion capture datasets of Human3.6M [33] and HumanEva-I [34] using two standard evaluation protocols.

4.1.1 DataSets

Human3.6M dataset is one of the most widely used dataset for 3D human pose estimation and it contains data captured through four synchronized cameras at 50 Hz. This dataset contains 3.6 million video frames for recording 11 professional actors performing 15 different actions including sitting down, sitting, purchasing, eating, talking on phone, etc. HumanEva-I is a much smaller dataset and it contains data captured through three cameras at different views at 60 Hz. HumanEva-I contains 7 calibrated video sequence, which are synchronized with 3D body poses obtained from a motion capture system. HumanEva-I contains video frames of 4 subjects performing 6 common actions such as walking, jogging, gesturing.

4.1.2 Evaluation protocols

In our experiments, we use two evaluation protocols as in previous studies [14, 30, 35]. Protocol#1 uses the mean per-joint position error (MPJPE) performance metric, which represents the mean Euclidean distance between the estimated result of an algorithm and the corresponding ground truth. Protocol#2 applies a procrustes analysis with the ground truth as a pre-processing step in MPJPE calculation. Performance metric of protocol #2 is abbreviated

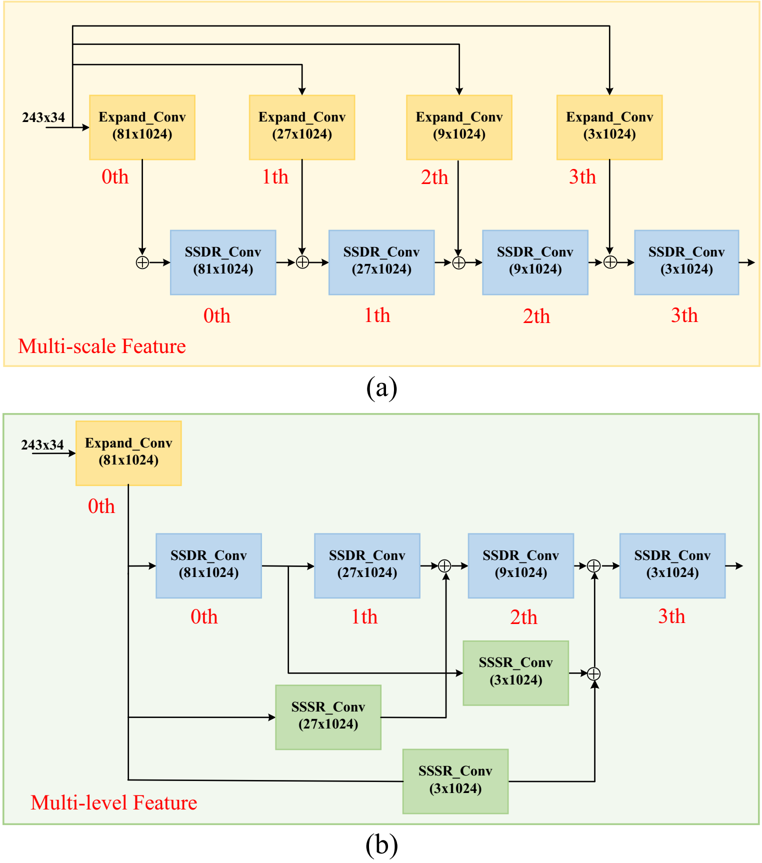

Fig. 7 Multi-scale and multi-level features. (**a**) The multi-scale feature branch network;(**b**) The multi-level feature branch network

to P-MPJPE. These evaluation protocols compute the distance errors between the predictions and the ground truth of the joints' positions.

4.2 Implementation details

The proposed model is implemented in PyTorch framework on a GeForce RTX 3090 GPU. We train the model using the Ranger optimizer.

For Human3.6M, we set the batch size $b = 1024$, set the initial learning rate to 1e-3, apply a shrinkage factor $a = 0.95$ after each epoch, and the model is trained for 80 epochs. As in many related works [14, 35], we adopt a 17-joint human skeleton compatible with 2D human pose estimation. We employ action sequences S1, S5, S6, S7, S8 for training and S9, S11 for testing. For HumanEva-I, we set $b = 128$ and $a = 0.995$ and train for 1000 epochs. We train and test with video sequences of "Walk" and "Jog" actions performed by subjects S1, S2 and S3 in the HumanEva-I dataset. We build three models with T=27, T=81, and T=243, where T denotes the size of the receptive field.

Table 1 Quantitative evaluation results of MPJPE (mm) on Human3.6M. CPN detections of 2D key points are used as input

Method	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Tekin et al. [39]	54.2	61.4	60.2	61.2	79.4	78.3	63.1	81.6	70.1	107.3	69.3	70.3	74.3	51.8	63.2	69.7
Martinez et al. [11]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Hossain et al. [12]	48.4	50.7	57.2	55.2	63.1	72.6	53.0	51.7	66.1	80.9	59.0	57.3	62.4	46.6	49.6	58.3
Lee et al. [17]	**40.2**	49.2	47.8	52.6	50.1	75.0	50.2	43.0	**55.8**	73.9	54.1	55.6	58.2	43.3	43.3	52.8
Xu et al. [30]	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. [14]	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
Liu et al. [35]	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
Ours(T=243 CPN causal)	43.8	46.6	43.5	46.6	50.2	56.0	44.7	43.6	58.3	68.4	48.0	46.2	48.4	34.2	34.1	47.5
Ours(T=243 CPN)	42.4	**44.6**	**40.3**	**43.9**	**47.1**	**53.7**	**42.5**	**41.6**	55.9	**62.7**	**45.0**	**42.6**	46.0	31.8	**31.4**	**44.8**

Note: Best: bold and red; Second best: underlined and blue

Table 2 Quantitative evaluation results of MPJPE (mm) on Human3.6M. Ground truth 2D key points are used as input

Method	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Martinez et al.[11]	37.7	44.4	40.3	42.1	48.2	54.9	44.4	42.1	54.6	58.0	45.1	46.4	47.6	36.4	40.4	45.5
Hossain et al. [12]	35.2	40.8	37.2	37.4	43.2	44.0	38.9	35.6	42.3	44.6	39.7	39.7	40.2	32.8	35.5	39.2
Lee et al. [17]	**32.1**	36.6	34.4	37.8	44.5	49.9	40.9	36.2	44.1	45.6	35.3	35.9	37.6	30.3	35.5	38.4
Pavllo et al. [14]	35.8	40.2	32.7	35.7	38.5	45.5	40.6	36.1	48.8	47.3	37.8	39.7	38.7	27.8	29.5	37.8
Xu et al. [30]	35.8	38.1	**31.0**	35.3	35.8	43.2	37.3	**31.7**	**38.4**	45.5	35.4	36.7	36.8	27.9	30.7	35.8
Liu et al. [35]	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
Ours(T=243 GT)	33.9	**34.9**	31.2	**32.8**	**31.4**	**35.7**	**36.9**	32.2	39.6	42.2	**32.3**	**33.5**	**31.9**	**25.2**	**25.9**	**33.3**

Note: Best: bold and red; Second best: underlined and blue

Table 3 Quantitative evaluation results of P-MPJPE(mm) on Human3.6M. CPN detections of 2D keypoints are used as input

Method	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT	Avg.
Martinez et al.[11]	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
Hossain et al. [12]	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
Pavllo et al. [14]	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. [35]	**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
Ours(T=243 CPN causal)	34.7	37.2	35.5	38.5	38.6	43.9	34.8	33.4	46.8	54.4	38.6	35.1	38.1	26.9	27.8	37.6
Ours(T=243 CPN)	33.5	35.6	**32.7**	36.4	36.3	41.8	**33.2**	**32.1**	**44.4**	**50.1**	**36.5**	33.4	**35.9**	**25.1**	**25.9**	**35.5**
Ours(T=243 GT)	24.3	27.1	23.4	25.1	23.9	27.5	26.9	23.9	30.2	33.7	25.0	25.0	25.2	19.5	20.0	25.4

Note: Best: bold and red; Second best: underlined and blue

4.3 Comparison with state of the art methods

We compare the performance of the proposed method with some state-of-the-art methods.

As in some previous work [14, 30, 35, 36], we use two types of 2D joint data for performance evaluation on Human3.6M: the 2D key points detected from Cascaded Pyramid Network (CPN) [37] and the ground truth 2D key points. The results of the experiments of MPJPE are presented in Tables 1 and 2, and those of M-MPJPE are provided in Table 3. Our method achieves competitive performance on Human3.6M under MPJPE and P-MPJPE. When using ground truth 2D keypoints, our method achieves an error reduction of approximately 11% compared to TCN [14]. Some qualitative comparisons of our method and TCN are shown in Fig. 8.

To further investigate the performance of our method and TCN, we compare the prediction accuracy between our method and TCN with various receptive fields. As shown in Figs. 9 and 10, our method achieves a smaller estimation error and a shorter runtime. In addition, we compare the prediction accuracy for the actions between our method and TCN. As shown in Table 4, our approach performs better than TCN for all actions, especially for the three heavily occluded actions of sitting down, sitting, and purchasing. There is a 3.1mm reduction in error for sitting down in the performance metric of MPJPE. The results confirm that our approach can mitigate the issues of depth ambiguity and self-occlusion.

The duration of a video in the HumanEva-I dataset is usually much shorter than that in the Human3.6M dataset, so We evaluate our method with a receptive field size of 27 under P-MPJPE. As in previous works [14], we adopt Mask R-CNN [38] to obtain 2D poses as input. The empirical results of our method on HumanEva-I are presented in Table 5.

Finally, we compared the complexity of our model with other methods, as shown in Table 6. The comparison mainly focuses on the number of parameters in the models and the expected floating-point operations, specifically matrix multiplication. As seen in the table, liu et al. [35]'s method has slightly fewer model parameters than ours, but their floating-point calculation times are considerably higher. Overall, our model demonstrates lower complexity.

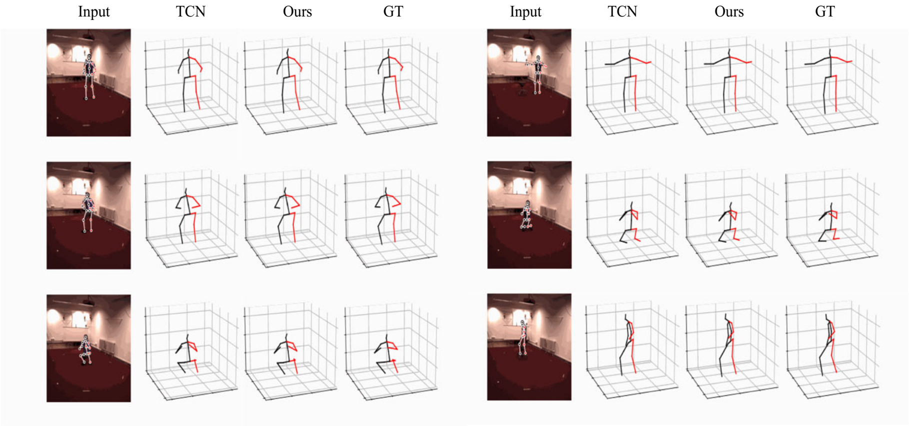

Fig. 8 Qualitative results of our method and TCN on Human3.6M [33]

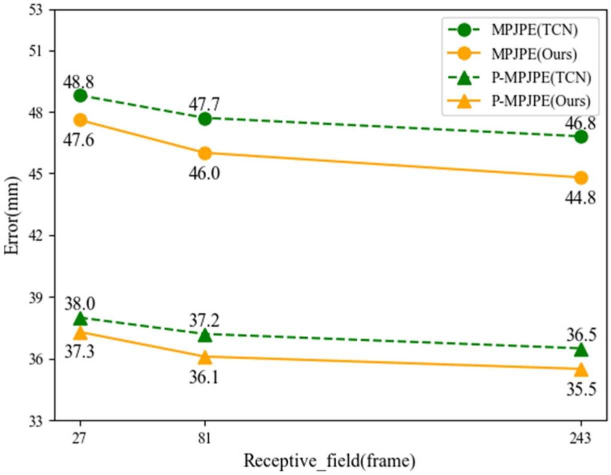

Fig. 9 Comparison with TCN with different receptive fields on Human3.6M

4.4 Ablation studies

In our ablation study, we use 2D pose predictions of CPN on the Human3.6M dataset for evaluating the impact of each component in our network.

4.4.1 Spatial-wise Separable Residual Convolution

Our spatial-wise separable residual convolution contains three modifications: (1) the order of depth-wise convolution and point-wise convolution are swapped; (2) a residual connection is inserted; (3) the Mish activation function is used to replace the RelU activation function. These three modifications are applied independently to reveal the effect of each modification.

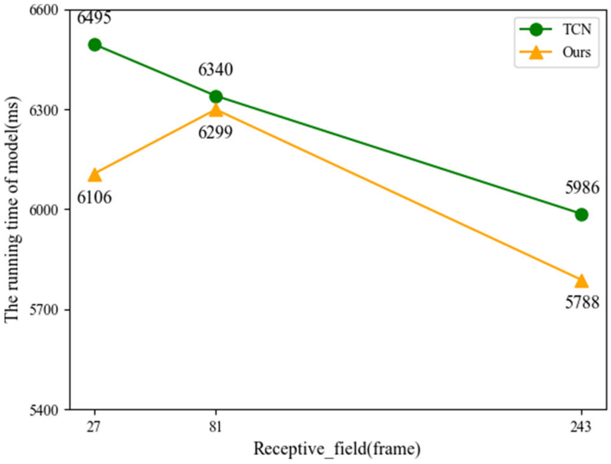

Fig. 10 Comparison of the running time of the TCN model with different receptive fields

Table 4 Quantitative evaluation results on Human3.6M between our method and TCN[14] under MPJPE (mm) and P-MPJPE (mm). CPN detections 2D key points are used as input

Protocols	Method	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD	Walk	WalkT	Avg.
MPJPE	Pavllo et al. [14]	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
	Ours (T=243 CPN)	**42.4**	**44.6**	**40.3**	**43.9**	**47.1**	**53.7**	**42.5**	**41.6**	**55.9**	**62.7**	**45.0**	**42.6**	**46.0**	**31.8**	**31.4**	**44.8**
P-MPJPE	Pavllo et al. [14]	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
	Ours (T=243 CPN)	**33.5**	**35.6**	**32.7**	**36.4**	**36.3**	**41.8**	**33.2**	**32.1**	**44.4**	**50.1**	**36.5**	**33.4**	**35.9**	**25.1**	**25.9**	**35.5**

Note: Best: bold and red; Second best: underlined and blue

Table 5 Quantitative evaluation results on HumanEva-I under P-MPJPE (mm). Mask R-CNN detections of 2D key points are used as input

Method	Walk			Jog			Avg.
	S1	S2	S3	S1	S2	S3	
Martinez et al. [11]	19.7	17.4	46.8	26.9	18.2	18.6	24.6
Pavlakos et al. [40]	18.8	12.7	29.2	23.5	15.4	14.5	19.0
Lee et al. [17]	18.6	19.9	30.5	25.7	16.8	17.7	21.5
Pavllo et al. [14]	13.9	10.2	46.6	20.9	13.1	13.8	19.8
Ours (T=27 Mask R-CNN)	**13.5**	**10.0**	**27.4**	**17.3**	**12.3**	**12.9**	**15.6**
Ours (T=27 GT)	8.0	6.7	11.6	10.3	7.7	8.9	8.9

Note: Best: bold and red; Second best: underlined and blue

The results are shown in Table 7. Performance progressively improves as the convolutional structure changes.

4.4.2 Effect of activation function

We investigate the effectiveness of activation functions as shown in Table 8. MPJPE achieved with Mish function is 1.2 mm lower than ReLU on our baseline model with receptive field size T=27.

4.4.3 Effect of the number of channels

We study how the number of channels C affects the performance of the baseline model. As shown in Figure 11, the error gradually decreases as the number of channels increases. For channel values between 128 and 512, MPJPE decreases significantly. When the number of channels is larger than 512, the MPJPE reduction is significantly smaller. When the number of channels increases from 1024 to 2048, the error is reduced by up to 1.7%. However, model parameters grow exponentially and training is significantly slower. Hence, we set the number of channels to 1024 in the following experiments.

Table 6 Computational complexity of various models under Protocol#1 trained on ground-truth 2D poses

Model	Parameters	≈FLOPs
Hossain et al.[12]	16.96M	33.88M
Liu et al. (T=243)[35]	14.52M	97.78M
Lee et al. [17]	31.00M	—
Pallo et al. (T=27)[14]	8.56M	17.09M
Pallo et al. (T=81)	12.75M	25.48M
Pallo et al. (T=243)	16.96M	33.87M
Ours (T=27)	4.48M	8.92M
Ours (T=81)	10.90M	17.53M
Ours (T=243)	15.22M	30.34M

Table 7 Ablation study on the spatial-wise separable residual convolution structure under MPJPE (mm) and P-MPJPE (mm)

Method	Params	MPJPE	↓	P-MPJPE	↓
The Standard Convolution	16.95M	46.8	-	36.5	-
The Original Structure	8.58M	49.0	-	38.6	-
+Modify The Order of Execution	8.58M	46.8	2.2	36.9	2.0
+Residual Connection	8.58M	47.4	1.6	37.3	1.3
+Activation Function (Mish)	8.58M	47.7	1.3	37.7	0.9
Our Network (T=243 CPN)	8.58M	45.9	3.1	36.4	2.2

Table 8 Ablation study on feature networks in our method under MPJPE (mm) and P-MPJPE(mm)

Activation Function	Params	MPJPE	↓	P-MPJPE	↓
ReLU	4.37M	49.5	-	38.5	-
Mish	4.37M	48.3	1.2	37.4	1.1

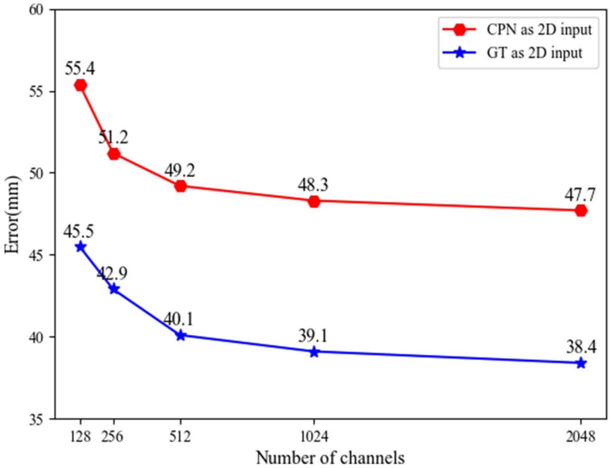

Fig. 11 Ablation studies on a different number of channels under MPJPE(mm). CPN: CPN detection of 2D key points, and GT: ground truth of 2D key points

Table 9 Ablation study on feature networks in our method under MPJPE (mm) and P-MPJPE(mm)

Model Method	T = 27		T = 81		T = 243	
Baseline (CPN)	48.3	37.4	46.9	36.8	45.9	36.4
+Multi-scale (CPN)	47.6	37.3	46.4	36.6	45.4	36.1
+Multi-level (CPN)	-	-	46.3	36.2	45.3	35.6
Our (CPN)	47.6	37.3	46.0	36.1	44.8	35.5
Our (GT)	37.0	27.9	35.5	26.1	33.3	25.4

4.4.4 Effect of multi-scale and multi-level features

To evaluate the impact of multi-scale and multi-level features, we test the performance of our network model with and without the multi-scale feature branch network or multi-level feature branch network.

The results in Table 9 show that, introducing multi-scale and multi-level features into our model with receptive field size 243 reduces the error by 1.1 mm.

5 Conclusions

In this work, we use multi-scale and multi-level features to build the powerful feature extraction capability and fuse three branch networks to better mitigate depth ambiguity and resolve self-occlusion. Quantitative results demonstrate that our approach can effectively improve the accuracy of 3D pose estimation. In future work, we will focus on more efficient architecture to extract more advanced features.

Acknowledgements This work was jointly supported by Fujian University-Industry Cooperation Project of Research and Industrialization of Machine Vision driven Automatic sorting System under Grant 2021H6030, National Natural Science Foundation of China under Grants 61972166 and 62372190. In addition, we would like to thank Xiaoxiao Wu for her work on the comparison of model complexity, as well as the revision to the manuscript.

Author Contributions Liling Fan and Zhenguo Gao conceived and designed the study; Liling Fan and Kunliang Jiang performed the experiments; Liling Fan, Zhenguo Gao and Weixue Zhou analyzed the data; and Liling Fan , Zhenguo Gao, Yanmin Luo, Kunliang Jiang and Weixue Zhou wrote the paper with input from all authors. All authors read and approved the final manuscript.

Funding This work was jointly supported by Fujian University-Industry Cooperation Project of Research and Industrialization of Machine Vision driven Automatic sorting System under Grant 2021H6030, National Natural Science Foundation of China under Grants 61972166 and 62372190.

Availability of data and materials The data and materials that support the findings of this study are available from the corresponding author upon reasonable request. The Human3.6m and HumanEva datasets were used in this study. The human3.6m dataset can be obtained from the official website: http://vision.imar.ro/human3.6m/description.php. The humaneva dataset is available for non-commercial research purposes and can be downloaded from: http://humaneva.is.tue.mpg.de/.

Code Availability Code is available at: https://github.com/fll123/3Dhumanpose.

Declarations

Conflict of interest/Competing interests 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.

References

1. Li Y, Ji B, Shi X, Zhang J, Kang B, Wang L (2020) TEA: Temporal excitation and aggregation for action recognition. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 906–915. https://doi.org/10.1109/CVPR42600.2020.00099
2. Xiu Y, Li J, Wang H, Fang Y, Lu C (2019) Pose flow: Efficient online pose tracking. In: British machine vision conference, (BMVC). https://doi.org/10.48550/arXiv.1802.00977

3. Amin S, Andriluka M, Rohrbach M, Schiele B (2013) Multiview pictorial structures for 3D human pose estimation. In: British machine vision conference(BMVC), vol 1. https://ias.in.tum.de/_media/spezial/bib/sikandar2013bmvc.pdf
4. Mandery C, Terlemez O, Do M, Vahrenkamp N, Asfour T (2015) The kit whole-body human motion database. In: International conference on advanced robotics (ICAR), pp 329–336. https://doi.org/10.1109/ICAR.2015.7251476
5. Henry P, Krainin M, Herbst EV, Ren X, Fox D (2014) RGB-D mapping: Using depth cameras for dense 3D modeling of indoor environments. In: Experimental robotics: The 12th international symposium on experimental robotics (ISER), pp 477–491. https://doi.org/10.1007/978-3-642-28572-1_33
6. Neverova N, Wolf C, Taylor GW, Nebout F (2014) Multiscale deep learning for gesture detection and localization. In: European conference on computer vision workshops (ECCV), pp 474–490. https://doi.org/10.1007/978-3-319-16178-5_33
7. Toshev A, Szegedy C (2014) Deeppose: Human pose estimation via deep neural networks. In: Conference on computer vision and pattern recognition (CVPR), pp 1653–1660. https://doi.org/10.1109/cvpr.2014.214
8. Nie X, Feng J, Zhang J, Yan S (2019) Single-stage multi-person pose machines. In: IEEE/CVF international conference on computer vision (ICCV), pp 6950–6959. https://doi.org/10.1109/ICCV.2019.00705
9. Wang Z, Nie X, Qu X, Chen Y, Liu S (2022) Distribution-aware single-stage models for multi-person 3D pose estimation. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 13086–13095. https://doi.org/10.1109/CVPR52688.2022.01275
10. Wu Q, Xu G, Li M, Chen L, Zhang X, Xie J (2018) Human pose estimation method based on single depth image. IET Comput Vis 12(6):919–924. https://doi.org/10.1049/iet-cvi.2017.0536
11. Martinez J, Hossain R, Romero J, Little JJ (2017) A simple yet effective baseline for 3D human pose estimation. In: IEEE international conference on computer vision (ICCV), pp 2659–2668. https://doi.org/10.1109/ICCV.2017.288
12. Hossain MRI, Little JJ (2018) Exploiting temporal information for 3D human pose estimation. In: Proceedings of the european conference on computer vision (ECCV), pp 68–84. https://doi.org/10.1007/978-3-030-01249-6_5
13. Wang L, Chen Y, Guo Z, Qian K, Lin M, Li H, Ren JS (2019) Generalizing monocular 3D human pose estimation in the wild. In: IEEE/CVF international conference on computer vision workshop (ICCVW), pp 4024–4033. https://doi.org/10.1109/ICCVW.2019.00497
14. Pavllo D, Feichtenhofer C, Grangier D, Auli M (2019) 3D human pose estimation in video with temporal convolutions and semi-supervised training. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 7745–7754. https://doi.org/10.1109/CVPR.2019.00794
15. Pavlakos G, Zhu L, Zhou X, Daniilidis K (2018) Learning to estimate 3D human pose and shape from a single color image. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 459–468. https://doi.org/10.1109/CVPR.2018.00055
16. Pavlakos G, Zhou X, Derpanis KG, Daniilidis K (2017) Coarse-to-fine volumetric prediction for single-image 3D human pose. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 1263–1272. https://doi.org/10.1109/CVPR.2017.139
17. Lee K, Lee I, Lee S (2018) Propagating LSTM: 3D pose estimation based on joint interdependency. In: Proceedings of the european conference on computer vision (ECCV), pp 123–141. https://doi.org/10.1007/978-3-030-01234-2_8
18. Zhao L, Peng X, Tian Y, Kapadia M, Metaxas DN (2019) Semantic graph convolutional networks for 3D human pose regression. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 3420–3430. https://doi.org/10.1109/CVPR.2019.00354
19. Sifre L, Mallat S (2014) Rigid-motion scattering for texture classification. Comput Sci 3559:501–515. https://doi.org/10.48550/arXiv.1403.1687
20. Rangnekar A, Mokashi N, Ientilucci EJ, Kanan C, Hoffman MJ (2020) AeroRIT: A new scene for hyperspectral image analysis. IEEE Trans Geosci Remote Sens 58(11):8116–8124. https://doi.org/10.1109/tgrs.2020.2987199
21. Hara K, Kataoka H, Satoh Y (2018) Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 6546–6555. https://doi.org/10.1109/CVPR.2018.00685
22. Yan S, Xiong Y, Lin D (2018) Spatial temporal graph convolutional networks for skeleton-based action recognition. 32(1). https://doi.org/10.48550/arXiv.1801.07455
23. Tome D, Russell C, Agapito L (2017) Lifting from the deep: Convolutional 3D pose estimation from a single image. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 5689–5698. https://doi.org/10.1109/CVPR.2017.603

24. Yang S, Wen J, Fan J (2022) Ghost shuffle lightweight pose network with effective feature representation and learning for human pose estimation. IET Comput Vis 16(6):525–540. https://doi.org/10.1049/cvi2.12110
25. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 770–778. https://doi.org/10.1109/CVPR.2016.90
26. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C (2018) MobileNetV2: Inverted residuals and linear bottlenecks. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 4510–4520. https://doi.org/10.1109/CVPR.2018.00474
27. Li S, Chan AB (2015) 3D human pose estimation from monocular images with deep convolutional neural network. In: Asian conference on computer vision (ACCV), pp 332–347. https://doi.org/10.1007/978-3-319-16808-1_23
28. Glorot X, Bordes A, Bengio Y (2011) Deep sparse rectifier neural networks. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics (AISTATS), vol 15, pp 315–323. https://proceedings.mlr.press/v15/glorot11a.html
29. Zhao Q, Sheng T, Wang Y, Tang Z, Chen Y, Cai L, Ling H (2019) M2Det: A single-shot object detector based on multi-level feature pyramid network. Proceedings of the AAAI conference on artificial intelligence. 33:9259–9266. https://doi.org/10.1609/aaai.v33i01.33019259
30. Xu T, Takano W (2021) Graph stacked hourglass networks for 3D human pose estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 16100–16109. https://doi.org/10.1109/CVPR46437.2021.01584
31. Wu Y, Gao J (2021) Multi-scale spatial-temporal transformer for 3D human pose estimation. In: 5th International conference on vision, image and signal processing (ICVISP), pp 242–247. https://doi.org/10.1109/ICVISP54630.2021.00051
32. Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. In: Proceedings of the european conference on computer vision (ECCV), pp 483–499. https://doi.org/10.48550/arXiv.1603.06937
33. Ionescu C, Papava D, Olaru V, Sminchisescu C (2014) 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. https://doi.org/10.1109/TPAMI.2013.248
34. Sigal L, Balan AO, Black MJ (2010) Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. Int J Comput Vis 87:4–27. https://doi.org/10.1007/s11263-009-0273-6
35. Liu R, Shen J, Wang H, Chen C, Cheung S-c, Asari V (2020) Attention mechanism exploits temporal contexts: Real-time 3D human pose reconstruction. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 5063–5072. https://doi.org/10.1109/CVPR42600.2020.00511
36. Zou Z, Tang W (2021) Modulated graph convolutional network for 3D human pose estimation. In: IEEE/CVF international conference on computer vision (ICCV), pp 11457–11467. https://doi.org/10.1109/ICCV48922.2021.01128
37. Chen Y, Wang Z, Peng Y, Zhang Z, Yu G, Sun J (2018) Cascaded pyramid network for multi-person pose estimation. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 7103–7112. https://doi.org/10.1109/CVPR.2018.00742
38. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask R-CNN. In: IEEE international conference on computer vision (ICCV), pp 2980–2988. https://doi.org/10.1109/ICCV.2017.322
39. Tekin B, Márquez-Neila P, Salzmann M, Fua P (2017) Learning to fuse 2D and 3D image cues for monocular body pose estimation. In: IEEE international conference on computer vision (ICCV), pp 3961–3970. https://doi.org/10.1109/ICCV.2017.425
40. Pavlakos G, Zhou X, Daniilidis K (2018) Ordinal depth supervision for 3D human pose estimation. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 7307–7316. https://doi.org/10.1109/CVPR.2018.00763

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.