JointFusionNet: Parallel Learning Human Structural Local and Global Joint Features for 3D Human Pose Estimation

Zhiwei Yuan[1], Yaping Yan[1], Songlin Du[1(✉)], and Takeshi Ikenaga[2]

[1] Southeast University, Nanjing 210096, China
sdu@seu.edu.cn
[2] Waseda University, Kitakyushu 808-0135, Japan

Abstract. 3D human pose estimation plays important roles in various human-machine interactive applications, but how to efficiently utilize the joint structural global and local features of human pose in deep-learning-based methods has always been a challenge. In this paper, we propose a parallel structural global and local joint features fusion network based on inspiring observation pattern of human pose. To be specific, it is observed that there are common similar global features and local features in human pose cross actions. Therefore, we design global-local capture modules separately to capture features and finally fuse them. The proposed parallel global and local joint features fusion network, entitled JointFusionNet, significantly improve state-of-the-art models on both intra-scenario H36M and cross-scenario 3DPW datasets and lead to appreciable improvements in poses with more similar local features. Notably, it yields an overall improvement of 3.4 mm in MPJPE (relative 6.8% improvement) over the previous best feature fusion based method [22] on H36M dataset in 3D human pose estimation.

Keywords: 3D human pose estimation · Human structural joint features · Feature fusion

1 Introduction

Human pose estimation (HPE), aiming to build human body representation (such as body skeleton and body shape), is a longstanding computer vision problem. 3D Human pose estimation has been applied to numerous applications, including motion recognition and analysis, human-computer interaction, virtual reality (VR), security identification and so on. However, this task is extremely challenging due to 1) the deep ambiguity in 2D-to-3D space transformation where a 2D keypoint corresponds to multiple 3D poses, 2) the insufficiency of labeled dataset due to the high cost of obtaining labels and 3) self-occlusion of human pose. Thanks to the success of deep neural networks, the generalization performance of the deep-learning-based method has risen sharply [13,16,25], making it have a broader prospect.

E. Pimenidis et al. (Eds.): ICANN 2022, LNCS 13532, pp. 113–125, 2022.
https://doi.org/10.1007/978-3-031-15937-4_10

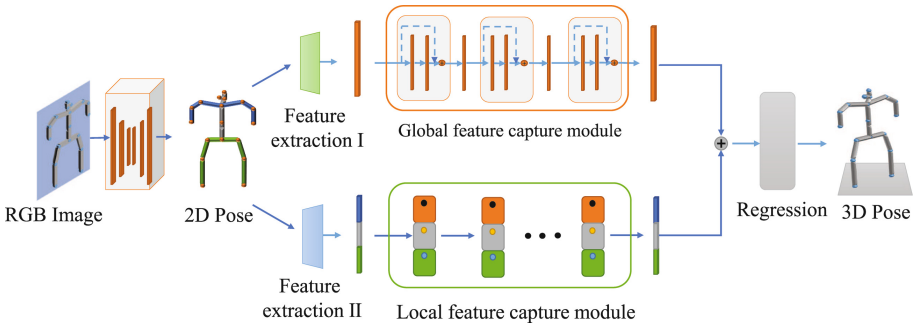


Fig. 1. Architecture of our proposed global-local features fusion network for 3D human pose estimation. Parallel full-connection module and group-connection module capture the global and local features of human pose respectively.

Improving the generalization performance of deep-learning-based 3D human pose estimation models is still a challenging problem. Surprisingly, we found some inspiring observations in the existing 3D human pose estimator: the prediction errors of keypoints of the human body have a high correlation with the structure of human pose, as shown in Fig. 2. Different human pose cross actions have high similarities in local features, and there are some samples as shown in Fig. 2(c). These inspiring observations indicate that there are communal global features in human pose cross actions, and regular similar local features existing in pose of different action categories. Whether it is possible to design a model based on this inspiring pattern for 3D human pose estimation.

To realize this idea, we designed a parallel fusion network where full-connection network and group-connection network learn to capture global-feature and local-feature of human pose, respectively. In this work, full-connection network connect all the input features and output features indiscriminately to learn the global information. Group-connection network connect the input features and output features in the group with global information representation to focus on learning the local information of human pose. Based on this motivation, a parallel fusion network which learn human structural local and global joint features (JointFusionNet) is designed for 3D human pose estimation, as shown in Fig. 1.

In extensive comparisons to state-of-the-art techniques, JointFusionNet exhibits performance considerably on 3D human pose estimation. More importantly, experiments show that JointFusionNet can not only outperform previous work, but elevate huge performance on pose with more similar local features, such as Sit, Greet and Phone. Moreover, various ablation studies validate our proposed approach. The main contributions are summarized as follows:

- The methods of global and local features capture modules are proposed respectively to model human pose based on inspiring observation.
- A network structure that fuses the global and local joint features of human pose is designed to improve the estimation performance.

– Extensive comparisons and various ablation studies to validate our proposed JointFusionNet for single-frame 3D human pose estimation.

2 Related Works

Extensive research has been conducted on 3D human pose estimation and global-local features fusion. In the following, we briefly review methods that are related to our approach.

2.1 3D Human Pose Estimation

Existing deep-learning-based 3D pose estimation methods mainly follow two frameworks: end-to-end methods and two-stage methods. End-to-end methods regress 3D pose directly from the input image [15], which is extremely expensive to acquire labeled datasets, although those methods avoid error accumulation in two stages. Thanks to the high accuracy of 2D pose estimators, the two-stage method has become a major popular solution for 3D human pose estimation. The two-stage methods [13,25,26] first employ off-the-shelf 2D pose estimators to extract 2D pose from the input image and then establish the mapping from 2D pose to 3D pose. Simultaneously, considering the sequence information in video, Pavllo *et al.* [16] proposed a network that combines multi-frame sequence information for pose estimation. The methods of multi-view fusion [5] are also applied to 3D human pose estimation due to the natural existence of multiple-view cameras or sensors in the dataset or reality. Furthermore, the transform network based on the attention mechanism [10] is also used to mine spatial and temporal information in 3D human pose estimation. Simultaneously, lacking diversity in existing labeled 3D human pose dataset restricts the generalization ability of deep learning based methods. Therefore, Li *et al.* [9] proposed a method to synthesize massive paired 2D-3D human skeletons with evolution strategy. JointPose [21] further jointly performs pose network estimation and data augmentation by designing a reward/penalty strategy for effective joint training. In this paper, we focus on the universal transformation from 2D pose to 3D pose in the two-stage method, as much as possible to fuse the global and local features of human pose at the same time.

2.2 Global-Local Features Fusion

The method of considering global and local features has long been applied to deep learning models, such as part-based branching network [17] for 2D human pose estimation. Martinez *et al.* [13] proposed a simple yet effective full connection network to learn the mapping relationship of 2D-3D space, in which the keypoints are fully connected but do not pay attention to local connection features. Based on the connection relationship of the graph model, Ma *et al.* [11] proposed a pose estimation model considering context node information, which does not consider local groups based on human pose structure information. Zeng *et al.* [22] proposed a grouping and reorganization pose estimation model based on the local

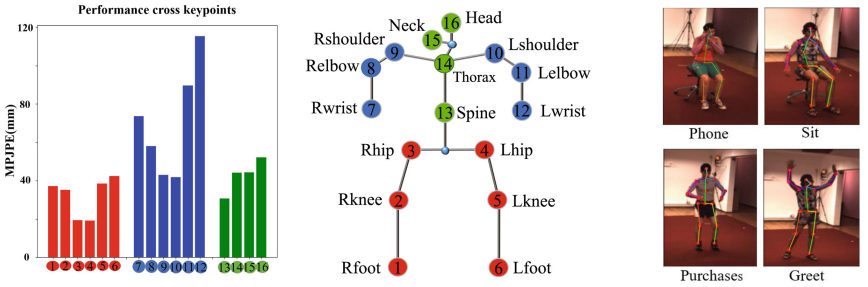
group of human structure information, which does not fully consider the global information of human pose. In this paper, we consider the global and local joint features of the human pose in parallel, hence propose JointFusionNet that fuses global and local features for 3D human pose estimation.

3 Method

In this section, we propose global- and local-feature capture module to learn the observed patterns of human pose and design a parallel fusion network for 3D human pose estimation.

3.1 Inspiring Pattern of Human Pose

The inspiring and regular observed patterns in human pose be applied to 3D human pose estimation models. Pattern I is the estimation performance pattern cross keypoints in existing 3D human pose estimator [22]: the estimation performance cross keypoints are not only quite different, but also show a regular pattern, as shown in Fig. 2(a) and Fig. 2(b). Pattern II is the similar local features of human pose: different human pose cross actions have specific similarities in local features. Here are some examples in H36M, as shown in Fig. 2(c).



(a) Keypoints performance (b) Keypoints of human pose (c) Example pose in H36M

Fig. 2. Inspiring observation pattern in human pose.

Pattern of estimation performance cross keypoints reveals that the performance of keypoints has a great correlation with the structure of human pose. The closer the keypoints to the center of the human body (like hip, shoulder and spine), the smaller the estimation error; the further away from the part (like foot, wrist and head), the greater the error compared to a structurally adjacent keypoint. The keypoints with low error represent global features of human pose, which are related to all the keypoints of human pose; The keypoints with high error represent local features of human pose, which are related to only part of keypoints. Based on Pattern I, also considering the partitioning of human pose used in [14], the keypoints of human pose are divided into groups, for example, 3 groups in different colors in Fig. 2(a) and Fig. 2(b).

Pattern of similar local features of human pose reveals that different human pose cross actions have high similarities in local features. Considering the structure and kinematics of human pose, the local features are limited and appear repeatedly in different actions. For example, the local feature of Sitting in the lower body appears in the action categories of Phone; the local feature of standing in the lower body appears in action of Purchases, Greet and so on. Further, the group of local features are not strongly related to each other, for example, the posture of the arm and of the legs are not highly correlated.

3.2 Global and Local Features Fusion

Inspired by the observed pattern of estimation error cross keypoints and similar local features of human pose, we explore this pattern to design the architecture of 3D pose estimation network, and proposed JointFusionNet, as shown in Fig. 1. In JointFusionNet, the 2D pose of the RGB image is estimated by the existing 2D human pose estimator, then full connection module and group connection module capture the global and local features of human pose. Finally, these features are fused and regressed to 3D pose.

We propose the global-local feature capture module, as shown in Fig. 3, to learn to capture global feature information and local feature information of human pose. Then a parallel structure is used to design the fusion network that fuses the learned features.

(a) Full-connection layer network (FCN) (b) Group-connection layer network (GCN)

Fig. 3. Global-local features capture module

Global feature capture module is a full-connection layer network (FCN) [13], learning the global features of human pose when processing the encoding vector representing the global information of the human body, as shown in Fig. 3(a). It can be noted that in the global feature capture module, each output features and each intermediate feature is connected to all of the input features indiscriminately, allowing it to learn the global information represented by each feature. Simultaneously, residual connections [4] are used as a technique to improve generalization performance.

Local feature capture module is a group-connection layer network (GCN) with Low-Dimensional Global Context (LDGC) [22], learning the local features of human pose when processing the encoding vector representing the group local

information of the human body, as shown in Fig. 3(b). According to the observed patterns in the human body and previous researches [14,22], we divide the keypoints into groups, which are used to capture the local features of human pose. And Low-Dimensional Global Context is used to learn to represent the relationship between local features and the whole pose.

Given the keypoints of 2D human pose $X = \{X_i|i, ..., N\} \in \mathbb{R}^{2N}$, where N is the number of keypoints. Formally, the global feature of human pose can be expressed as

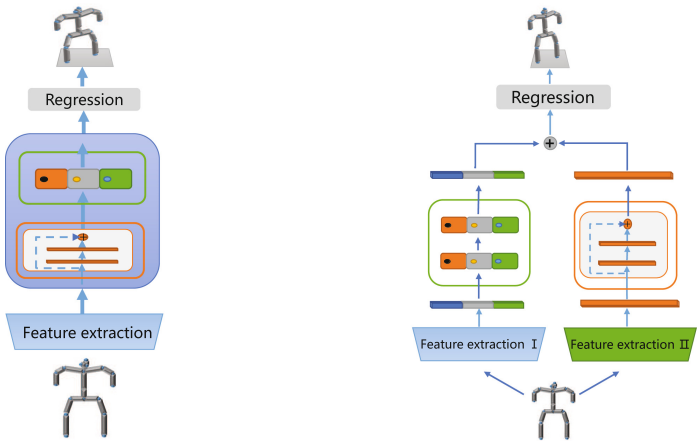
$$F_{global} = FCN(X). \tag{1}$$

Then the keypoints can be divided into k groups $X^k = \left\{X_i^k|i, ..., N_k\right\} \in \mathrm{R}^{2N}$, where k represents the number of groups, and N_k represents the number of keypoints in k^{th} group. The local feature of human pose can be expressed as

$$F_{local}^k = GCN^k(X^k), \tag{2}$$

where F_{local}^k represents local feature in k^{th} group.

Arrangement of Feature Capture Modules. Full connection module and group connection module capture the global and local features of human pose respectively. When the representations of global features and local features are determined, how to fuse these two features becomes a key issue. At this time, the global-feature capture module and the local-feature capture module can be placed in a parallel or sequential manner, as shown in Fig. 4. Based on the previous feature fusion based method [22] and our research experiments, the parallel arrangement gives a better result than a sequential arrangement, which means learning the information of local features and global features separately and then conducting the fusion of the two.



(a) Sequential manner (b) Parallel manner

Fig. 4. Conceptual difference arrangement of full connection module and group connection module.

Table 1. The MPJPE (mm) of the SOTA methods on the H36M dataset under protocol #1 and protocol #2, respectively. Best performance is marked with bold font. Dim: representation dimension.

Method authors		Performance															
Protocol #1		Dire	Disc	Eat	Greet	Phone	Photo	Pose	Purch	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkP	**Avg**
Zhou *et al.* [26]	ICCV'17	54.8	60.7	58.2	71.4	62.0	65.5	53.8	55.6	75.2	111.6	64.2	66.1	51.4	63.2	55.3	64.9
Martinez *et al.* [13]	ICCV'17	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
Pavlakos *et al.* [16]	CVPR'18	48.5	54.4	54.4	52.0	59.4	65.3	49.9	52.9	65.8	71.1	56.6	52.9	60.9	44.7	47.8	56.2
Yang *et al.* [20]	CVPR'18	51.5	58.9	50.4	57.0	62.1	65.4	49.8	52.7	69.2	85.2	57.4	58.4	43.6	60.1	47.7	58.6
Zhao *et al.* [25]	CVPR'19	47.3	60.7	51.4	60.5	61.1	**49.9**	47.3	68.1	86.2	**55.0**	67.8	61.0	42.1	60.6	45.3	57.6
Iskakov *et al.* [7]	ICCV'19	41.9	49.2	46.9	47.6	50.7	57.9	41.2	50.9	57.3	74.9	48.6	44.3	**41.3**	52.8	42.7	49.9
Wang *et al.* [18]	ICCV'19	44.7	48.9	47.0	49.0	56.4	67.7	48.7	47.0	63.0	78.1	51.1	50.1	54.5	40.1	43.0	52.6
Ci *et al.* (LCN) [2]	ICCV'19	46.8	52.3	44.7	50.4	52.9	68.9	49.6	46.4	60.2	78.9	51.2	50.0	54.8	40.4	43.3	52.7
Pavllo *et al.* [16]	CVPR'19	47.1	50.6	49.0	51.8	53.6	61.4	49.4	47.4	59.3	67.4	52.4	49.5	55.3	39.5	42.7	51.8
Cai *et al.* [1]	ICCV'19	46.5	48.8	47.6	50.9	52.9	61.3	48.3	45.8	59.2	64.4	51.2	48.4	53.5	39.2	41.2	50.6
Zeng *et al.* [22]	ECCV'20	44.5	48.2	47.1	47.8	51.2	56.8	50.1	45.6	59.9	66.4	52.1	45.3	54.2	39.1	40.3	49.9
Li *et al.* [9]	CVPR'20	45.6	**44.6**	49.3	49.3	52.5	58.5	46.4	44.3	53.8	67.5	49.4	46.1	52.5	41.4	44.4	49.7
Xu *et al.* [19]	CVPR'20	40.6	47.1	45.7	46.6	50.7	63.1	**45.0**	47.7	56.3	63.9	49.4	46.5	51.9	38.1	42.3	49.2
Gong *et al.* [3]	CVPR'21	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	50.2
Zeng *et al.* [23]	ICCV'21	43.1	50.4	**43.9**	45.3	46.1	57.0	46.3	47.6	56.3	61.5	47.7	47.4	53.5	35.4	**37.3**	47.9
Zhan *et al.* [24]	CVPR'22	44.7	48.7	48.7	48.4	51.0	59.9	46.8	46.9	58.7	61.7	50.2	46.4	51.5	38.6	41.8	49.7
Ours (Dim: 2048)		38.7	46.7	50.5	40.6	45.1	62.9	47.2	40.0	43.0	76.7	47.3	42.0	48.3	35.9	41.9	47.1
Ours (Dim: 4096)		**37.3**	47.9	49.6	**38.8**	**43.7**	62.6	45.9	**37.9**	**41.6**	80.4	**45.6**	**41.7**	51.2	**33.6**	39.6	**46.5**
Protocol #2		Dire	Disc	Eat	Greet	Phone	Photo	Pose	Purch	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkP	**Avg**
Martinez *et al.* [13]	ICCV'17	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
Pavlakos *et al.* [16]	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
Yang *et al.* [20]	CVPR'18	**26.9**	30.9	**36.3**	39.9	43.9	47.4	**28.8**	**29.4**	36.9	58.4	41.5	30.5	**29.5**	42.5	32.2	37.7
Wang *et al.* [18]	ICCV'19	33.6	38.1	37.6	38.5	43.4	48.8	36.0	35.7	51.1	63.1	41.0	38.6	40.9	30.3	34.1	40.7
Ci *et al.* (LCN) [2]	ICCV'19	36.9	41.6	38.0	41.0	41.9	51.1	38.2	37.6	49.1	62.1	43.1	39.9	43.5	32.2	37.0	42.2
Pavllo *et al.* [16]	CVPR'19	36.0	38.7	38.0	41.7	40.1	45.9	37.1	35.4	46.8	53.4	41.4	36.9	43.1	30.3	34.8	40.0
Cai *et al.* [1]	ICCV'19	36.8	38.7	38.2	41.7	40.7	46.8	37.9	35.6	47.6	51.7	41.3	36.8	42.7	31.0	34.7	40.2
sXu *et al.* [19]	CVPR'20	33.6	37.4	37.0	37.6	39.2	46.4	34.3	35.4	45.1	52.1	40.1	35.5	42.1	29.8	35.3	38.9
Li *et al.* [9]	CVPR'20	34.2	**34.6**	37.3	39.3	38.5	45.6	34.5	32.7	40.5	**51.3**	37.7	35.4	39.9	29.9	34.5	37.7
Gong *et al.* [3]	CVPR'21	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	39.1
Ours (Dim: 2048)		28.8	40.0	40.0	34.4	49.2	**33.2**	32.0	31.4	32.4	60.1	37.7	28.8	41.0	28.5	33.9	36.8
Ours (Dim: 4096)		27.4	39.1	38.3	**33.4**	**31.5**	47.3	31.0	30.1	**31.5**	63.5	**36.0**	**27.6**	42.5	**26.9**	**31.9**	**35.9**

4 Experiments

In this section, we quantitatively evaluate the effectiveness of JointFusionNet and visualize the observed patterns and further explain the performance of JointFusionNet cross actions. The ablation study analyzes the effects of global and local features, representation dimension, and grouping strategy.

4.1 Datasets, Evaluation Metrics and Details

Human3.6M [6] is a large benchmark widely used for 3D human pose estimation with 11 professional actors collected by the motion sensor. Following conventional works, data from 5 actors (subject 1, 5, 6, 7, 8) are used for training, and data from other 2 actors (subject 9, 11) are used for testing. We use MPJPE and PA-MPJPE for evaluation.

3DPW [12] is the first dataset in the wild with more complicated motions and scenes for 3D human pose estimation evaluation. To verify generalization of the proposed method, we use its test set for evaluation with MPJPE and PA-MPJPE as metric.

Evaluation Metrics. Following convention, we use the mean per joint position error (MPJPE) [6] for evaluation, as follows

$$MPJPE = \frac{1}{N}\sum_{i=1}^{N} \|J_i - J_i^*\|_2, \tag{3}$$

where N is the number of all joints, J_i and J_i^* are respectively the groundtruth position and the estimated position of the ith joint. Protocol #1 was directly calculated, Protocol #2 (Procrustes Analysis MPJPE, PA-MPJPE) was calculated after rigid transformation.

Implementation Details. To train the 3D human pose estimation network, we adopt Adam optimizer [8] with a learning rate initialized as 0.001 and decays at the rate of 0.95 after each epoch. We train JointFusionNet model for 60 epoches in PyTorch framework on NVIDIA RTX 2080 Ti GPU.

4.2 Comparison with State-of-the-Art Methods

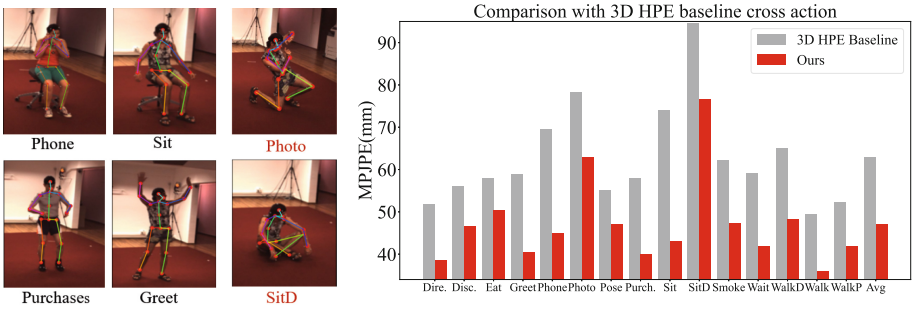
In this setting, we use the 2D pose detected by off-the-shelf 2D pose estimator as input of JointFusionNet, and set the representation dimension to 4096, grouping strategy to 5. We first compare our proposed method with the state-of-the-art methods using the standard subject protocol under Protocol #1 and Protocol #2. Table 1 shows that JointFusionNet yields an overall improvement over state-of-the-art methods, indicating strong generalization ability for 3D human pose estimation.

4.3 Cross-dataset Results on 3DPW

In this setting, we examine cross-dataset generalization ability of JointFusionNet by training the model on the Human3.6M training set and evaluating on the 3DPW test set. The performance of JointFusionNet is generally outperforming than that of previous work by a large margin. Notably, it yields an overall improvement of 14.8 mm (relative 13.8% improvement) over the previous best method [2] on 3DPW dataset. As shown in Table 2, proposed approach achieves the best cross-data generalization performance.

Table 2. Performance on the 3DPW test set

Method authors		Performance	
		Protocol #1	Protocol #2
Martinez *et al.* (FCN) [13]	ICCV'17	159.8	113.3
Pavllo *et al.* (1-frame) [16]	CVPR'19	–	146.3
Zhao *et al.* (SemGCN) (1-frame) [25]	CVPR'19	–	152.3
Ci *et al.* (LCN) [2]	ICCV'19	191.5	107.6
Cai *et al.* (ST-GCN) (1-frame) [1]	ICCV'19	–	154.3
Zeng *et al.* [22]	ECCV'20	169.0	110.7
Gong *et al.* (PoseAug) (1-frame) [3]	CVPR'21	–	130.3
Ours		**123.1**	**92.8**



(a) Example pose in H36M (b) Comparison with 3D HPE Baseline

Fig. 5. A visualization of example pose with huge and light improvement compared with 3D HPE Baseline cross actions.

4.4 Visualization and Explanation

This section visualizes some example human pose with local similar features in H36M and explains the performance comparison between JointFusionNet and 3D HPE Baseline method cross actions, as shown in Fig. 5. The local feature of sitting in the lower body and standing in the lower body appear in different action categories similarly. Correspondingly, the estimation performance of action (Such as Sit and Phone) that has more similar local features has a huge improvement (relative 26.4%, 36.8% improvement over the 3D HPE Baseline method [13]). On the contrary, there is still an improvement in the performance of actions with relatively less similar local features (Such as Photo and SitD), though it is difficult for JointFusionNet to fully learn the relationship between global and local features of the action with few local similar features.

4.5 Ablation Study

Effect of Global and Local Features. In proposed JointFusionNet, the global-feature capture module and local-feature capture module focu on learning global and local features of human pose respectively. Therefore, this set of experiments explores the role of the global features and local features separately. This experiments use global-feature capture module, local-feature capture module and parallel global-local-feature capture module to capture features, respectively. Compared to capturing the global or local features individually, the proposed global-local features fusion network is more efficient, as shown in Table 3.

Table 3. Performance under capturing different features

Representation features	Global	Local	Global and Local
Protocol #1	62.9	49.9	**46.5**
Protocol #2	47.7	38.7	**35.9**

Effect of Representation Dimension. In the Global-feature capture module, we use a high-dimensional feature representation of human pose. In this experiment, we set different representations to explore the effect of representation dimension. Higher-dimensional features represent the potential to learn to capture more complex interconnections, although inevitably pose challenges to network training, as shown in Table 4.

Table 4. Performance under different representation dimensions

Representation dimensions	1024	2048	4096	5120
Protocol #1	47.80	47.12	**46.50**	46.89
Protocol #2	37.67	36.75	**35.86**	36.61

Effect of Grouping Strategy. We compare the results of using different numbers of local groups in Table 5. Although there can be more and complex grouping strategies, we only set 3 commonly used strategies to explore the effect of grouping strategy. The way of grouping reveals the structural information of the human body. It is shown that the performance is best when the grouping method is consistent with intuitive perception, which indicates that a strong physical relationship among joints in a group is a prerequisite for learning effective local features.

Table 5. Performance under different grouping strategies

Group	5	3	2
Protocol #1	46.50	**46.25**	46.53
Protocol #2	**35.86**	35.88	36.16

5 Conclusion

In this paper, we proposed JointFusionNet, a structural global and local joint features fusion approach based on inspiring observation patterns, which improves generalization performance in 3D human pose estimation. The key idea is to design a parallel fusion network that captures global-features and local-features for more effective learning. Experimental results and ablation studies show that JointFusionNet outperforms state-of-the-art techniques, especially for poses with more similar local features.

Acknowledgment. This work was jointly supported by the National Natural Science Foundation of China under grant 62001110, the Natural Science Foundation of Jiangsu Province under grant BK20200353, and the "Zhishan Young Scholar" Program of Southeast University.

References

1. Cai, Y., et al.: Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2272–2281 (2019). https://doi.org/10.1109/ICCV.2019.00236
2. Ci, H., Wang, C., Ma, X., Wang, Y.: Optimizing network structure for 3D human pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2262–2271 (2019). https://doi.org/10.1109/ICCV.2019.00235
3. Gong, K., Zhang, J., Feng, J.: PoseAug: a differentiable pose augmentation framework for 3D human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8575–8584 (2021)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
5. He, Y., Yan, R., Fragkiadaki, K., Yu, S.I.: Epipolar transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7776–7785 (2020). https://doi.org/10.1109/CVPR42600.2020.00780
6. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: 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 (2014). https://doi.org/10.1109/TPAMI.2013.248
7. Iskakov, K., Burkov, E., Lempitsky, V., Malkov, Y.: Learnable triangulation of human pose. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7717–7726 (2019). https://doi.org/10.1109/ICCV.2019.00781

8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Li, S., Ke, L., Pratama, K., Tai, Y.W., Tang, C.K., Cheng, K.T.: Cascaded deep monocular 3D human pose estimation with evolutionary training data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6173–6183 (2020)
10. Li, W., Liu, H., Tang, H., Wang, P., Van Gool, L.: MHFormer: Multi-hypothesis transformer for 3D human pose estimation. arXiv preprint arXiv:2111.12707 (2021)
11. Ma, X., Su, J., Wang, C., Ci, H., Wang, Y.: Context modeling in 3D human pose estimation: a unified perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6234–6243 (2021). https://doi.org/10.1109/CVPR46437.2021.00617
12. von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 601–617 (2018)
13. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3D human pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2659–2668 (2017). https://doi.org/10.1109/ICCV.2017.288
14. Park, S., Kwak, N.: 3D human pose estimation with relational networks. CoRR abs/1805.08961 (2018). http://arxiv.org/abs/1805.08961
15. Pavlakos, G., Zhou, X., Daniilidis, K.: Ordinal depth supervision for 3D human pose estimation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7307–7316 (2018). https://doi.org/10.1109/CVPR.2018.00763
16. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3D human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7745–7754 (2019). https://doi.org/10.1109/CVPR.2019.00794
17. Tang, W., Wu, Y.: Does learning specific features for related parts help human pose estimation? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1107–1116 (2019). https://doi.org/10.1109/CVPR.2019.00120
18. Wang, J., Huang, S., Wang, X., Tao, D.: Not all parts are created equal: 3D pose estimation by modeling bi-directional dependencies of body parts. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7771–7780 (2019)
19. Xu, J., Yu, Z., Ni, B., Yang, J., Yang, X., Zhang, W.: Deep kinematics analysis for monocular 3D human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 896–905 (2020). https://doi.org/10.1109/CVPR42600.2020.00098
20. Yang, W., Ouyang, W., Wang, X., Ren, J., Li, H., Wang, X.: 3D human pose estimation in the wild by adversarial learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5255–5264 (2018).https://doi.org/10.1109/CVPR.2018.00551
21. Yuan, Z., Du, S.: JointPose: jointly optimizing evolutionary data augmentation and prediction neural network for 3D human pose estimation. In: Farkaš, I., Masulli, P., Otte, S., Wermter, S. (eds.) ICANN 2021, Part III. LNCS, vol. 12893, pp. 368–379. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86365-4_30

22. Zeng, A., Sun, X., Huang, F., Liu, M., Xu, Q., Lin, S.: SRNet: improving generalization in 3D human pose estimation with a split-and-recombine approach. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 507–523 (2020)
23. Zeng, A., Sun, X., Yang, L., Zhao, N., Liu, M., Xu, Q.: Learning skeletal graph neural networks for hard 3D pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 11436–11445 (2021)
24. Zhan, Y., Li, F., Weng, R., Choi, W.: Ray3D: ray-based 3D human pose estimation for monocular absolute 3D localization (2022). https://doi.org/10.48550/ARXIV.2203.11471, http://arxiv.org/abs/2203.11471
25. Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N.: Semantic graph convolutional networks for 3D human pose regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3425–3435 (2019)
26. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3D human pose estimation in the wild: a weakly-supervised approach. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 398–407 (2017). https://doi.org/10.1109/ICCV.2017.51