



Staged cascaded network for monocular 3D human pose estimation

Bing-kun Gao¹ · Zhong-xin Zhang¹ · Cui-na Wu¹ · Chen-lei Wu¹ · Hong-bo Bi¹

Accepted: 15 March 2022 / Published online: 23 April 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

The study of deep end-to-end representation learning for 2D to 3D monocular human pose estimation is a common yet challenging task in computer vision. However, current methods still face the problem that the recognized 3D key points are inconsistent with the actual joint positions. The strategy that trains 2D to 3D networks using 3D human poses with corresponding 2D projections to solve this problem is effective. On this basis, we build a cascaded monocular 3D human pose estimation network, which uses a hierarchical supervision network, and uses the proposed composite residual module (CRM) and enhanced fusion module (EFM) as the main components. In the cascaded network, CRMs are stacked to form cascaded modules. Compared with the traditional residual module, the proposed CRM expands the information flow channels. In addition, the proposed EFM is alternately placed with cascaded modules, which addresses the problems of reduced accuracy and low robustness caused by multi-level cascade. We test the proposed network on the standard benchmark Human3.6M dataset and MPI-INF-3DHP dataset. We compare the results under the fully-supervised methods with six algorithms and the results under the weakly-supervised methods with five algorithms. We use the mean per joint position error (MPJPE) in millimeters as the evaluation index and get the best results.

Keywords 3D human pose estimation · Residual network · Deep learning · Cascaded network

1 Introduction

The research of human pose estimation has aroused widespread concern in the computer vision field. The task aims to detect the location of human key points in pictures and videos to generate human poses and provide technological support for some tasks, such as action recognition [20, 45], human-computer interaction [15], autonomous driving [6], etc. The proposal of deep neural networks (DNNs) has greatly improved the network performance of computer vision. At the same time, some excellent networks have also appeared, such as VGG [45], GoogLeNet [48], ResNet [17], DenseNet [18], and so on. In this case, many effective human pose estimation networks have been proposed by scholars, such as stacked hourglass networks [33], pyramid residual networks [12], and high-resolution networks [46].

Compared with 2D human pose estimation, 3D human pose estimation can provide more comprehensive spatial position information. However, the characteristics of images captured by monocular cameras make it difficult for 3D networks to estimate the accurate pose from the RGB image [29]. The lack of depth information and the change of camera viewpoint and appearance [13] have led to poor performance of 3D human pose [8, 28, 39, 40]. For this task, two mainstream architectures are usually adopted: one-stage methods [16, 45, 49, 53, 57] and two-stage methods [32, 41, 55]. The former maps the pixel intensity to the 3D pose directly, while the latter is achieved in two stages [5, 13, 44, 52, 55]. In the first stage, the algorithm locates 2D human key points based on the appearance information of the image; in the second stage, the network model uses a 2D network to process the image and then uses geometric information to generate a 3D skeleton from the 2D joints [9, 18].

Since the two-stage method is based on 2D human pose estimation, there are a lot of research results that can be used for reference. Therefore, more scholars tend to use the two-stage methods to complete the task of 3D human pose estimation. The 2D to 3D network is the focus of our research in the second stage [34]. Aiming at the problem

✉ Hong-bo Bi
bhbdq@126.com

Extended author information available on the last page of the article.

that most current networks cannot be used for direct training from 2D to 3D, we build a cascaded deep network based on the research of the evolutionary dataset proposed in [23] to generate 3D poses. Our main contributions are as follows:

1. We propose a monocular cascaded deep network. We compare the results under the fully-supervised methods with six algorithms and the results under the weakly-supervised methods with five algorithms and get the best result using MPJPE as the evaluation index.
2. We propose a composite residual module, which provides a richer flow of information, and expands the channel from one to four.
3. We propose an EFM, which adopts the parallel fusion structure of CRM. We add a fusion module at the junction of adjacent cascaded modules. EFM integrates the previous cascaded module's output and initially processes the original information to provide fusion information with different depths for the subsequent cascaded modules.

2 Related work

The single RGB image estimation methods are roughly divided into generative and discriminative methods in monocular 3D human pose estimation. The parameterized model is processed in the traditional generation method to fit into the image observation algorithm. These algorithms use PCA models [2, 37], graphical models [4, 7], or deformed meshes [19, 22, 37] to represent 3D human poses. Discriminative methods [1, 5] directly learn the mapping from image observation to 3D pose. Among them, deep neural networks (DNNs) use discriminative methods, and this task's completion generally adopts one-stage or two-stage methods (Fig. 1).

2.1 One-stage human pose estimation

The one-stage methods can directly estimate the 3D human pose, which can make full use of the end-to-end training advantage of the CNN network [50] and save time consumption. However, one-stage methods lack sufficient label data and excellent deep networks.

The one-stage method proposed by Nie et al. [36] has a simple framework and can be directly extended from 2D images to 3D images for human pose estimation and achieves a detection rate that is not inferior to the two-stage methods. The adversarial learning framework is proposed to deal with the complex environment in the wild [19, 53]. Instead of defining hard-coded rules to constrain the pose estimation results, the model extracts the 3D human pose structure learned from the fully annotated data set into the 2D pose annotated field image. The research of Ikhsanul Habibie et al. is also applicable to the wild environment [16]. The network includes a new disentangled hidden space encoding explicit 2D and 3D features and uses a newly learned projection model from the predicted 3D pose for supervision. Georgios Pavlakos [38] proposes a coarse-to-fine prediction scheme for the initial estimation. The model solves the problems caused by increasing large dimensions and realizes the iterative refinement and repeated processing of image features.

2.2 Two-stage human pose estimation

Compared with one-stage methods, two-stage methods have first-mover advantages due to the 2D human pose estimation research foundation. The 2D human pose estimation algorithms are extended or processed to complete the 3D pose generation, saving a lot of scientific research resources and time costs. Generally, the first stage of the two-stage methods uses a 2D landmark detector to process 2D images. The second stage is a cascaded 3D coordinate regression model used to generate a 3D pose.

Among the two-stage 3D human pose estimation algorithms, researchers are more inclined to study the second-stage 2D to 3D algorithms 2D to 3D networks. Julieta Martinez et al. found an effective method that uses a relatively simple deep feedforward network to obtain experimental data about 30% higher than the best-reported results [27]. The temporal information is added to a relatively complex deep network, and the material information across the 2D joint position sequence is used to estimate the 3D pose sequence. They design a sequence-to-sequence network composed of hierarchically normalized LSTM units to impose temporal smoothness constraints during the training process [41]. Francesc Moreno-Noguer

Fig. 1 Comparison with the visual results of Li et al. [23] on the MPI-INF-3DHP dataset. We obtained more accurate recognition results (legs etc., for view 2)



Institute [32] uses the distance matrix regression method, which leverages a simple neural network to enhance the prediction matrix’s positivity and symmetry, thereby improving the accuracy of the joint point regression. In the regression from graph convolutional networks (GCNs) to 3D coordinates, semantic graph convolutional networks (SemGCN) [55] can be learned through end-to-end training without additional supervision or manual rules.

Currently, the cascaded models are heavily used in image processing, and these models have achieved better recognition accuracy [3, 11, 25, 46, 54, 56]. For example, Haoran Bai et al. [3] used a super-resolution cascaded network to process blurred images to obtain higher resolution images. Diba et al. [14] proposed a cascaded model using a fully convolutional network to delineate candidate regions and segment the image to accomplish object saliency detection. Liu et al. [25] proposed cascaded networks with sequential localization of nodes to force messages from joints of independent components, such as the head and torso, to distal joints, such as the wrist or ankle. In addition, in human pose estimation, representative cascaded pyramid networks [12] use a two-stage model: GlobalNet and RefineNet, enabling the extraction of joint points and refinement of joint points, respectively. In terms of 3D human pose estimation, the staged cascaded network provides significant improvements in accuracy, and Li et al. [23] implemented deeply trained 3D pose estimation based on an evolved dataset.

To solve the problem of 3D data’s scarcity, Helge Rhodin et al. [42] proposes to replace most of the annotations with multiple views only during training and present a joint estimation method of camera pose and human pose. In video-based 3D human pose estimation, time information is widely used. [40] uses temporal convolution and semi-supervised methods to prove that the fully convolutional

model based on dilated temporal convolution can effectively use 2D keypoint information to estimate 3D human pose from the video.

3 Framework

In this section, we briefly introduce the two-stage network and internal structure. The framework uses CRM as a basic block, stacks them to form a cascaded structure, and adds EFM to enhance transmission between modules. The extraction of 2D key points is the basis for 3D pose estimation; we hope that the 2D network can retain more high-resolution information. Therefore, we use HRNet [46] as the backbone network. The size of the heat map predicted by the original model is 96×72 ; the pixel shuffle super-resolution layer [44] is added to the model, and the size of the returned heat map is 384×288 .

We use *TAG-Net* [23] to represent the proposed two-stage model, as shown in Fig. 2. To obtain a geometric representation of the model, we use the following function to express the appearance of the image

$$\hat{P} = TAG(x) = G(A(x)) \tag{1}$$

We use x as the initial input RGB image. In the first stage, x passes through a high-resolution network to generate a 2D keypoint heat map, represented by $A(x)$. The number of regression key points is expressed as $k = 17$. We coordinate the 2D keypoints $c = (x_i, y_i)_{i=1}^k$. In the cascaded module, we perform geometric processing on 2D coordinates in stages and use logical operations to infer 3D keypoint coordinates $p = (x_i, y_i, z_i)_{i=1}^k$, where we use $G(c)$ (geometric stage) to Indicates the process (Table 1).

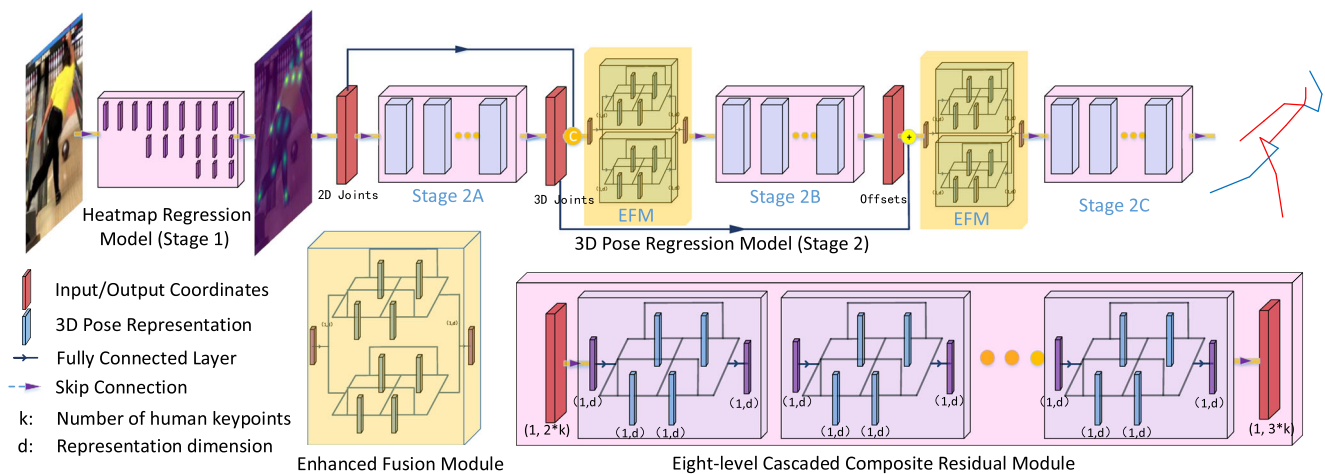


Fig. 2 Our cascaded 3D human pose estimation architecture. Above: a two-stage model. Bottom: Enhanced fusion module and 3D cascaded residual module

Table 1 Average 2D keypoint localization errors for H36M testing set in pixels. U: Heatmaps, upsampling.S: use soft-argmax

Backbone	Extension	#Params	FLOPs	Error
CPN [42]	-	-	13.9G	5.40
HRN [46]	-	63.6M	32.9G	4.98
HRN	+U	63.6M	32.9G	4.64
HRN	+U+S	63.6M	32.9G	4.36

New cascaded networks As shown in Fig. 2, our new cascaded human pose estimation architecture is divided into two stages. We use HRNet [46] to achieve accurate 2D coordinate detection in the first stage. In the second stage, we use different numbers of feedforward neural networks to stack as the main architecture of the cascaded 3D coordinate regression model. Our network consists of three cascaded modules, and each cascaded module uses eight layers of CRM to cascade (including a total of 24 CRMs). This structure ensures sufficient network depth and facilitates comparison with advanced methods.

Since the mapping from 2D coordinates to 3D joints is highly nonlinear and difficult to learn, we propose a cascaded 3D coordinate regression model as

$$\hat{P} = G(c) = \sum_{t=1}^T D_t(i_t, \theta_t) \quad (2)$$

where D_t is the t th deep learner in the cascade parametrized by θ_t and takes input i_t . As shown at the top of Fig. 2, the first learner D_1 in the cascade directly predicts 3D coordinates while the later ones predict the 3D refinement $\delta_p = (\delta_{x_i}, \delta_{y_i}, \delta_{z_i})_{i=1}^k$.

To train the heatmap regression model $A(x)$, we download training videos from the official website of H36M. We use the provided bounding box to crop the character and fill the cropped image with zeros to fix the aspect ratio to 4 : 3. Then we adjust the size of the supplied image to 384×288 . The size of the target heatmaps is equal

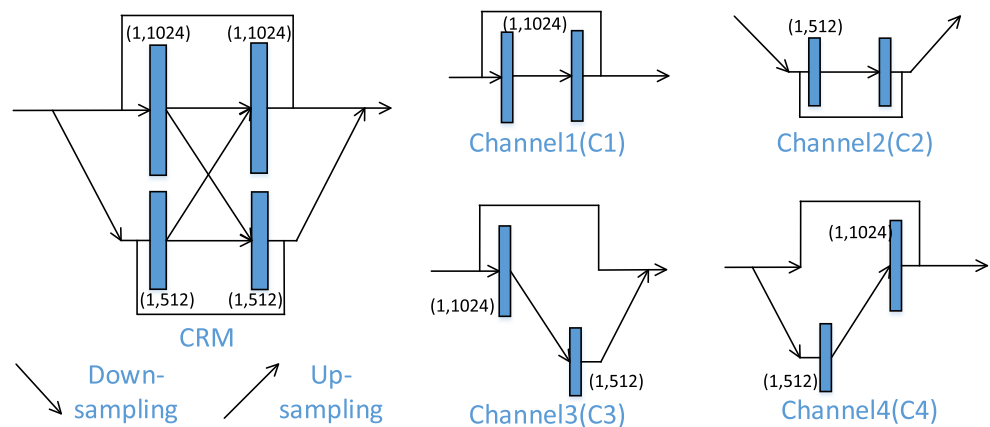
to the size of the input image, and a Gaussian dot is drawn for each human key point $\phi(x)_{i=1}^k$. The average value of Gaussian dots is the ground truth 2D keypoint position, and its standard deviation is 8 pixels. To train the cascaded 3D pose estimation model $G(c)$ on the H36M, we downloaded the preprocessed human skeleton published by the author of [34] in Github. Each deep learner in the cascaded model is trained with L_2 loss.

3.1 Composite residual module

The conventional research of 3D human pose estimation does not emphasize the selection and processing of the bottom network. The residual network tends to process characteristic information with simple forward feedback, making the processed information obtain low-level semantic information. The information flow of the residual block cannot be expanded well in the process of multiple cascades [12, 19, 28]. Therefore, we improve the classic residual network, and its main structure is shown in Fig. 3 (CRM). We propose the parallel double residual form as the basic residual block and add a fully connected layer to the parallel residual network. CRM enriches the information flow channel from a single channel to 4 channels. We combine the parallel residual network with the cascaded backbone network. This supervision method compensates for amplifying errors caused by the deepening of the cascaded network.

As shown in Fig. 3 (four channels of the CRM), the increase of channels makes the network produce two transmission modes (In Fig. 3, C1 and C2 represent the same mode, and C3 and C4 represent the same mode.), four channels of information flow channels (Channels C1, C2, C3, C4 in Fig. 3) [19]. The increase of channels expands the way of information transmission. The two-channel types (In Fig. 3, C1 and C2 represent the same mode, and C3 and C4 represent the same mode.) cause differences in the way and ability to process information, making the

Fig. 3 Illustrations of the four modes of the composite residual module. We design a parallel residual module with different sizes and add down-sampling and up-sampling to the module, producing four channels. C1: high-scale residual structure. C2: low-scale residual structure. C3: high-to-low-scale residual structure (downsampling). C4: low-to-high-scale residual structure (upsampling)



processed information richer more conducive to the use of global information. The cascaded network will continuously produce new features in information transmission. The use of CRM increases the types of useful information and maintains the robustness of the network.

In CRM, combining high and low scales residual blocks can obtain various types of information [16, 31, 35, 51]. In addition, although the cascade of multiple CRMs retains more helpful information, it also generates false information and amplifies useless information [19]. We use skip connections to receive lower-semantic information so that the network can simultaneously have semantic information of different depths [17, 37, 46].

When using CRM to generate 3D poses, we first map 2D coordinates to represent vectors with dimensions 1024 and 512, respectively. When processing vectors, CRM will transform the dimensions according to the four modes shown in Fig. 3. The diversity of dimensions and transformation forms can effectively improve the learning ability of deep models [3, 11, 12]. In the cascaded module, we cascaded eight CRMs. In each CRM, we add batch normalization [56] and dropout (dropout rate 0.5) [12, 32, 33, 49, 54]. Hossain and Little [41] used gradient descent to train the cascade network sequentially. Despite the fact that CRM increases the number of parameters in the cascaded network, we found that the cascaded model was robust to predictions of 3D coordinates, which is also common in 2D models [23].

3.2 Enhanced fusion module

We propose an EFM and alternately place it with cascaded modules to form the cascaded network. The purpose is to merge the features of the previous layer with the features processed by the cascaded modules. At the same time, EFM can initially process the original information incorporated by the skip connection to supplement the original information.

The network uses a fully connected layer for transmission and uses a double-layer CRM as the main structure. The internal structure of the fusion module is shown at the bottom in Fig. 2. The fusion module supervises the cascaded network by introducing original information and plays a role in correcting the flow of information in multiple branches. We add connected channels to the CRM and complete the inter-layer fusion by adding a 3D pose representation at the beginning and the end. We test the residuals of one layer, two layers, and three layers. The results illustrate that the two-layer structure has the best network scale and accuracy performance.

In addition, we hope that the initial information is hierarchically integrated into the 3D network instead of directly fusing with in-depth information without

processing. EFM plays the role of step-by-step interaction between global details and detailed information. In EFM, the original data is divided into three forms: no CRM processing, single CRM processing, and two-level CRM processing (as shown in Fig. 2 stage2) to enhance different depths' fusion.

In the main network, information is processed sequentially by the cascaded modules, and many 3D vectors are mapped. Prior to 3D mapping via CRM, the initial information was in 2D coordinates. The information discrepancy makes the fusion information uncertain [23, 46], and EFMs progressively reduce this discrepancy. In erosion experiments, we found that adding two EFMs did not significantly increase the number of 3D network parameters.

4 Experiments

To validate our network's effectiveness, we train and test on the Human 3.6M dataset and its evolved dataset [23]. Our training used an NVIDIA 1080ti GPU, which took 10 hours. We use a training strategy of 0.001 when training 3k times and multiplying by 0.1 after every 3k times. To optimize each deep learner during the training process, we use the Adam optimizer with a learning rate of 0.001 in the cascaded network to train for 200 epochs. We use the mean per joint position error (MPJPE) measured in millimeters to evaluate the model's performance and test the three protocols under weakly-supervised, and fully-supervised methods [24, 28, 40, 43, 47, 55]. In both cases, the average position error of each joint is reduced.

4.1 Datasets and evaluation indicators

Human 3.6M (H36M) is one of the most extensive 3D human pose estimation dataset benchmarks. It has accurate 3D tags and has reference significance in 3D human pose estimation. We add the topic ID to S to represent the divided dataset. For example, S15 represents data from topics 1 and 5. Previous work revises the training data [20, 26, 28, 30, 53], and we use it as the initial population and evolve from it. We use MPJPE to measure in millimeters to evaluate the model's performance. We adopt two standard evaluation schemes. Protocol 1 is referred to as P1, which directly calculates the result of MPJPE. The difference between protocol P1* and P1 is that P1* uses ground truth 2D key points as input, eliminating the first-stage model's influence. Furthermore, we align the ground truth 3D key points with the predictions and evaluate Protocol 2 (P2).

MPI-INF-3DHP (3DHP) is a benchmark that we use to evaluate the generalization power of 2D-to-3D networks in invisible environments. We perform cross-dataset inference by providing the key points of 3DHP to G(c). Among them,

Table 2 Comparison with SOTA weakly-supervised methods. This paper reports the average MPJPE of all 15 actions for Human 3.6M under two protocols (P1 and P2). P1* refers to protocol 1 evaluated using ground truth 2D key points. The best performance is marked in bold

Method	P1	P1*	P2
Rhodin et al. (CVPR'18) [42]	-	-	66.4
Kocabas et al. (CVPR'19) [21]	65.3	-	57.2
Pavlo et al. (CVPR'19) [40]	64.7	-	-
Li et al. (ICCV'19) [24]	88.8	-	65.5
Li et al. (CVPR'20) [23]	62.9	50.5	47.5
Ours	60.1	45.0	44.4

the percentage of correct keypoints (PCK) indicates the accuracy of the 3D joint predictions under the specified threshold of the measurement [51]. At the same time, the area under the curve (AUC) is computed for a range of PCK thresholds.

4.2 Comparison with the state-of-the-arts

We compare the fully-supervised and weakly-supervised methods with previous models [10, 26]. To validate the superiority of the proposed network, we use the same data as [23, 28], use S1 as the initial input, conduct experiments under multiple protocols, and obtain corresponding data.

Weakly-supervised methods We take the weakly-supervised Methods as the research's focus. Since this method lacks sufficient 3D annotations to explore weakly-supervised human pose estimation methods, we refer to the methods of [21, 24, 40]. On this basis, we introduce the weakly-supervised method in [23] and use a small amount of dataset to simulate the lack of data to train our model. We compare the previous methods [21, 23, 24, 40, 42], and the results are shown in Table 2.

Table 3 Comparison with SOTA methods under fully-supervised setting [23, 27, 31, 43, 53, 55]. This paper reports the average MPJPE of all 15 actions for Human 3.6M under two protocols (P1 and P2). P1* refers to protocol 1 evaluated using ground truth 2D key-points. The best performance is marked in bold

Method	P1	P1*	P2
Martinez et al. (ICCV'17) [27]	62.9	45.5	47.7
Yang et al. (CVPR'18) [53]	58.6	-	37.7
Zhao et al. (CVPR'19) [55]	57.6	43.8	-
Sharma et al. (ICCV'19) [43]	58.0	-	40.9
Moon et al. (ICCV'19) [31]	54.4	35.2	-
Li et al. (CVPR'20) [23]	50.9	34.5	38.0
Ours	49.0	32.5	37.6

Table 4 The table illustrates the generalization results of the 3DHP dataset. MPJPE is evaluated without rigid transformation, and CE stands for cross-data set evaluation, and our model does not use training data in 3DHP. The best performance is marked in bold

Method	CE	PCK	AUC	MPJPE
Mehta et al.(ICCV'17) [28]	-	76.5	40.8	117.6
VNect(TOG'17) [30]	-	76.6	40.4	124.7
Zhou et al.(ICCV'17) [57]	-	69.2	32.5	137.1
Multi Person(3DV'18) [29]	-	75.2	37.8	122.2
OriNet(2018) [26]	-	81.8	45.2	89.4
Kanazawa(CVPR'18) [20]	CE	77.1	40.7	113.2
Yang et al.(CVPR'18) [53]	CE	69.0	32.0	-
Li et al.(CVPR'20) [23]	CE	81.2	46.1	99.7
Ours	CE	86.5	46.1	98.9

It is worth noting that the method of [21, 42] additionally adds multi-view consistency as additional supervision. In contrast, our experiment is an evaluation under a single perspective. In the weakly supervised method of [40], time information is used to assist the training of the model, while our approach does not have time convolution. In this case, we have compared five methods in total. When MPJPE is used as the evaluation index, the network model has the highest accuracy.

Fully-supervised methods We use the Human3.6M dataset for training. We use S15678 as the initial population, and Table 3 shows the performance comparison. In this case, our model also obtains competitive performance compared with other SOTA methods, which proves that our method is not limited to scarce data scenarios.

In the fully-supervised method, we compare with six algorithms of the average MPJPE of all 15 actions of H3.6M under three protocols (P1, P2, P1*). Compared with other SOTA methods, our model obtains a competitive

Table 5 Ablation study of Human 3.6M. The experiment evaluated model performance with mean per joint position error (MPJPE) measured in millimeters. Baseline: the original data [23]. +CRM: add CRM based on the baseline. +CRM+EFM: add CRM and EFM based on the baseline

Method	Training Data	P1	P1*
Problem Setting A: Weakly-supervised Learning			
Baseline	S1	62.9	50.5
+CRM	S1	61.7	47.0
+CRM+EFM(final)	Evolve(S1)	60.1	45.0
Problem Setting B: Fully-supervised Learning			
Baseline	S1	50.9	34.5
+CRM	S15678	50.1	33.2
+CRM+EFM(final)	Evolve(S15678)	49.0	32.5



Fig. 4 Cross-dataset inferences of $G(c)$ on MPI-INF-3DHP (The first and second columns are our results, and the third and fourth columns are the results of others [23])

performance. In the same experimental environment, the performance of P1, P2, and P1* is improved by 1.9, 2.0, and 0.4 respectively compared with [23]. The results show that our network can be applied to most scenarios.

In the experiment, we found that as the complexity of the data set increases, the training error also increases slightly, while the test error decreases somewhat. We use multiple subsets of H36M for pose estimation [19]. As the number of learners increases, the training's fallacy gradually decreases. The final result also indicate that our model does not overfit.

The limitation of the scene leads to the conclusion that the result of weakly-supervised methods is better than the fully-supervised methods. For the P1, P1*, and P2 protocols, the results of MPJPE have the same proportion of decline. The reason is that our 3D network optimizes the ability to process information [51]. When the influence of the first stage model exists and the ground-truth 3D poses are aligned with the predictions through rigid transformation, the error still drops by a considerable amount.

Cross-dataset generalization To verify the generalization ability of our network, we compare it with other methods on 3DHP (as shown in Table 4). In this experiment, we test on the 3DHP benchmark without using any training data of 3DHP, and the results proved that our network achieves competitive performance in the 3DHP benchmark. Compared with [19], our staged cascaded network effectively improves the model's generalization ability.

4.3 Ablation study

The ablation studies are conducted on H36M. We experiment separately with protocols S1 and S15678 under the fully-supervised and weak-supervised methods. To compare

with the previous results, we also test the dataset evolved by Li et al. [23]. The effect of using only CRM, using EFM and CRM, and the new cascaded network is verified. Table 5 summarizes the results (Fig. 4).

5 Conclusion

This paper proposes a new monocular cascaded network and uses the CRM to design the bottom network. At the same time, we add the enhanced fusion module to the cascaded network. The network model increases the transmission of practical information flow and improves the network's performance while ensuring that the overall scale of the network is controllable. We experiment and test on the Human 3.6M dataset, compared with other five related algorithms under weakly-supervised methods, and compared with different six related algorithms under fully-supervised methods, and obtain the best results, which proved the effectiveness of the network. To verify the generalization ability of our network, we conduct experiments on 3DHP and get advanced results.

References


1. Agarwal A, Triggs B (2005) Recovering 3d human pose from monocular images. *IEEE Trans Pattern Anal Mach Intell* 28(1):44–58
2. Akhter I, Black MJ (2015) Pose-conditioned joint angle limits for 3d human pose reconstruction. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 1446–1455
3. Bai H, Cheng S, Tang J, Pan J (2021) Learning a cascaded non-local residual network for super-resolving blurry images. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 223–232
4. Belagiannis V, Amin S, Andriluka M, Schiele B, Navab N, Ilic S (2014) 3d pictorial structures for multiple human pose estimation.

- In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1669–1676
5. Bo L, Sminchisescu C (2009) Structured output-associative regression. In: 2009 IEEE Conference on computer vision and pattern recognition. IEEE, pp 2403–2410
 6. Bogo F, Kanazawa A, Lassner C, Gehler P, Romero J, Black MJ (2016) Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: European conference on computer vision. Springer, pp 561–578
 7. Burenius M, Sullivan J, Carlsson S (2013) 3d pictorial structures for multiple view articulated pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3618–3625
 8. Chen W, Wang H, Li Y, Su H, Wang Z, Tu C, Lischinski D, Cohen-Or D, Chen B (2016) Synthesizing training images for boosting human 3d pose estimation. In: 2016 Fourth international conference on 3d vision (3DV). IEEE, pp 479–488
 9. Chen X, Yuille A (2014) Articulated pose estimation by a graphical model with image dependent pairwise relations. arXiv:1407.3399
 10. Chen X, Lin K-Y, Liu W, Qian C, Lin L (2019) Weakly-supervised discovery of geometry-aware representation for 3d human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10895–10904
 11. Chen X, Fu C, Zhao Y, Zheng F, Song J, Ji R, Yi Y (2020) Saliency-guided cascaded suppression network for person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3300–3310
 12. Chen Y, Wang Z, Peng Y, Zhang Z, Yu G, Sun J (2018) Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7103–7112
 13. Cheng Y, Bo Y, Bo W, Yan W, Tan RT (2019) Occlusion-aware networks for 3d human pose estimation in video. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 723–732
 14. Diba A, Sharma V, Pazandeh A, Pirsiavash H, Gool LV (2017) Weakly supervised cascaded convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 914–922
 15. Dix A, Finlay J, Abowd GD, Beale R (2000) Human-computer interaction Harlow ua
 16. Habibie I, Xu W, Mehta D, Pons-Moll G, Theobalt C (2019) In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10905–10914
 17. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
 18. Huang G, Liu Z, Maaten LVD, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
 19. Ji X, Qi F, Dong J, Shuai Q, Jiang W, Zhou X (2020) A survey on monocular 3d human pose estimation. *Virtual Real Intell Hardw* 2(6):471–500
 20. Kanazawa A, Black MJ, Jacobs DW, Malik J (2018) End-to-end recovery of human shape and pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7122–7131
 21. Kocabas M, Karagoz S, Akbas E (2019) Self-supervised learning of 3d human pose using multi-view geometry. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1077–1086
 22. Kolotouros N, Pavlakos G, Black MJ, Daniilidis K (2019) Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 2252–2261
 23. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6173–6183
 24. Li Z, Wang X, Wang F, Jiang P (2019) On boosting single-frame 3d human pose estimation via monocular videos. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 2192–2201
 25. Liu W, Chen J, Li C, Qian C, Chu X, Hu X (2018) A cascaded inception of inception network with attention modulated feature fusion for human pose estimation. In: Thirty-second AAAI conference on artificial intelligence
 26. Luo C, Chu X, Yuille A (2018) Orinet: A fully convolutional network for 3d human pose estimation. arXiv:1811.04989
 27. Martinez J, Hossain R, Romero J, Little JJ (2017) A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE international conference on computer vision, pp 2640–2649
 28. Mehta D, Rhodin H, Casas D, Fua P, Sotnychenko O, Weipeng Xu, Theobalt C (2017) Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 2017 International conference on 3d vision (3DV). IEEE, pp 506–516
 29. Mehta D, Sotnychenko O, Mueller F, Xu W, Sridhar S, Pons-Moll G, Theobalt C (2018) Single-shot multi-person 3d pose estimation from monocular rgb. In: 2018 International conference on 3d vision (3DV). IEEE, pp 120–130
 30. Mehta D, Sridhar S, Sotnychenko O, Rhodin H, Shafiei M, Seidel H-P, Xu W, Casas D, Theobalt C (2017) Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Trans Graph (TOG)* 36(4):1–14
 31. Moon G, Chang YJ, Lee KM (2019) Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10133–10142
 32. Moreno-Noguer F (2017) 3d human pose estimation from a single image via distance matrix regression. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2823–2832
 33. Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. In: European conference on computer vision. Springer, pp 483–499
 34. Nibali A, He Z, Morgan S, Prendergast L (2018) Numerical coordinate regression with convolutional neural networks. arXiv:1801.07372
 35. Nie Q, Liu Z, Liu Y (2020) Unsupervised 3d human pose representation with viewpoint and pose disentanglement. In: European conference on computer vision. Springer, pp 102–118
 36. Nie X, Feng J, Zhang J, Yan S (2019) Single-stage multi-person pose machines. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6951–6960
 37. Pavlakos G, Choutas V, Ghorbani N, Bolkart T, Osman AAA, Tzionas D, Black MJ (2019) Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10975–10985
 38. Pavlakos G, Zhou X, Derpanis KG, Daniilidis K (2017) Coarse-to-fine volumetric prediction for single-image 3d human pose.

- In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7025–7034
39. Pavlakos G, Zhou X, Derpanis KG, Daniilidis K (2017) Harvesting multiple views for marker-less 3d human pose annotations. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6988–6997
 40. Pavllo D, Feichtenhofer C, Grangier D, Auli M (2019) 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7753–7762
 41. Hossain MRI, Little JJ (2017) Exploiting temporal information for 3d pose estimation. arXiv:[arXiv-1711](https://arxiv.org/abs/1711.02222)
 42. Rhodin H, Spörri J, Katircioglu I, Constantin V, Meyer F, Müller E, Salzmann M, Fua P (2018) Learning monocular 3d human pose estimation from multi-view images. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8437–8446
 43. Sharma S, Varigonda PT, Bindal P, Sharma A, Jain A (2019) Monocular 3d human pose estimation by generation and ordinal ranking. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 2325–2334
 44. Shi W, Caballero J, Huszár F, Totz J, Aitken AP, Bishop R, Rueckert D, Wang Z (2016) Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1874–1883
 45. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:[1409.1556](https://arxiv.org/abs/1409.1556)
 46. Ke S, Xiao B, Liu D, Wang J (2019) Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5693–5703
 47. Sun X, Xiao B, Wei F, Liang S, Wei Y (2018) Integral human pose regression. In: Proceedings of the european conference on computer vision (ECCV), pp 529–545
 48. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
 49. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
 50. Tompson JJ, Jain A, LeCun Y, Bregler C (2014) Joint training of a convolutional network and a graphical model for human pose estimation. arXiv:[1406.2984](https://arxiv.org/abs/1406.2984)
 51. Wang J, Tan S, Zhen X, Xu S, Zheng F, He Z, Shao L (2021) Deep 3d human pose estimation: A review. *Computer Vision and Image Understanding*, p 103225
 52. Wu J, Xue T, Lim JJ, Tian Y, Tenenbaum JB, Torralba A, Freeman WT (2016) Single image 3d interpreter network. In: European conference on computer vision. Springer, pp 365–382
 53. Yang W, Ouyang W, Wang X, Ren J, Li H, Wang X (2018) 3d human pose estimation in the wild by adversarial learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5255–5264
 54. Yu D, Su K, Sun J, Wang C (2018) Multi-person pose estimation for pose tracking with enhanced cascaded pyramid network. In: Proceedings of the european conference on computer vision (ECCV) Workshops, pp 0–0
 55. Zhao L, Xi P, Yu T, Kapadia M, Metaxas DN (2019) Semantic graph convolutional networks for 3d human pose regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3425–3435
 56. Zhou T, Wang W, Qi S, Ling H, Shen J (2020) Cascaded human-object interaction recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4263–4272
 57. Zhou X, Huang Q, Sun X, Xue X, Wei Y (2017) Towards 3d human pose estimation in the wild: A weakly-supervised approach. In: Proceedings of the IEEE international conference on computer vision, pp 398–407

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

Affiliations

Bing-kun Gao¹ · Zhong-xin Zhang¹ · Cui-na Wu¹ · Chen-lei Wu¹ · Hong-bo Bi¹ 

Bing-kun Gao
bkao@126.com

Zhong-xin Zhang
zhangzhongxin2021@126.com

Cui-na Wu
cnw2021@126.com

Chen-lei Wu
wcl_master@126.com

¹ School of Electrical Information Engineering, Northeast Petroleum University, Daqing 163000, China