IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 16, 2019, accepted May 19, 2019, date of publication May 27, 2019, date of current version June 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2919154

Human Pose Estimation With Deeply Learned Multi-Scale Compositional Models

RUI WANG, ZHONGZHENG CAO, XIANGYANG WANG, ZHI LIU, AND XIAOQIANG ZHU

Key laboratory of Specialty Fiber Optics and Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, School of Communication and Information Engineering, Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China

Corresponding author: Xiangyang Wang (wangxiangyang@shu.edu.cn)

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 61771299, Grant 61771322, Grant 61375015, and Grant 61301027.

ABSTRACT Compositional models are meant for human pose estimation (HPE) due to their abilities to capture relationships among human body parts. Deeply learned compositional model (DLCM) utilizes deep neural networks to learn compositionality of human body parts and has achieved great improvements in human pose estimation. The DLCM has a hierarchical compositional architecture and bottom-up/top-down inference stages. The previous works have proven that multi-scale deep features are beneficial for computer vision tasks, such as classification and human body keypoints detection. However, learning multi-scale feature pyramids in DLCM has not been well explored. In this paper, we propose a new method to apply the multi-scale feature pyramid module to further improve the performance of the DLCM, which is named as deeply learned multi-scale compositional model (DLMSCM). We design multi-scale residual modules as the basic blocks to learn multi-scale deep features which can capture the scale variations of different body parts. With the multi-scale mechanism in the framework of the DLCM, the model can not only deal with scale variations of body parts but also find joints dependencies, therefore enforce the entire body joints structural constrains. As a result, more precise body keypoints detection can be acquired. Our approach outperforms the other state-of-the-art methods on two standard benchmarks datasets MPII and LSP for human pose estimation.

INDEX TERMS Compositional models, human pose estimation, multi-scale residual modules, scale variations, joints structural constrains.

I. INTRODUCTION

Human pose estimation (HPE) aims to predict the locations of body joints from a single image. It is a challenging problem due to the limited information of 2D images and the large variations in configuration and appearance of body parts. Early works often tackles the problem using graphical models [1] with hand crafted image features. Those methods often lack effective feature representations to characterize complex appearance variations of different people, so they are hard to be improved further.

Recently, many deep learning [2]–[7] based human pose estimation models have been proposed and their performances are improved continuously. Deep convolutional neural networks (CNNs) can learn rich feature representations directly from data.

The recent proposed approach, stacked hourglass network [2], achieves state-of-the-art performance without use of hand designed priors or graphical-model-style inference. The architecture supports repeated bottom-up, top-down inference across scales for large receptive field, so the model can capture relationships among body parts.

However, Hourglass networks still have some limitations. First, the model only learns features from one image scale. In many cases, in the same image, different human body parts may have different scales. So it is necessary to incorporate feature pyramid into hourglass networks to capture the characteristic of scale variations of body parts [3]. Second, when there exist self-occlusion among human body parts or overlapping with other people nearby, the model may predict implausible human pose. Because, in such situations, the model is trying to find similar features which might be in the background or belong to another person.

By contrast, human visions [8], [9] have the concepts of the structure and constraint of body parts, and can associate these with observed image features.

In this paper, as [10], we take advantage of Deeply Learned Compositional Model (DLCM) for human pose estimation.

The associate editor coordinating the review of this manuscript and approving it for publication was Shuhan Shen.

We use key points coding to represent a part and supervise its score map during the training step. This strategy compactly encodes the orientation, scale and shape of a part, in which to check the structural constraints of human body pose. The network is a 5-stacked Hourglass networks in which the basic blocks of the first three are the feature pyramid units.

The Deeply Learned Compositional Model (DLCM) with feature pyramid module strategy enables the model to correct implausible poses and urge the model to improve its prediction.

II. RELATED WORK

Recently, many methods [11], [12] have been developed by taking advantage of the deep Convolutional Neural Networks (CNNs). DeepPose [13] is the first deep learning based approach for human pose estimation, which takes the pose estimation as a body keypoints regression problem using convolutional Neural Networks, and outperforms previous classical approaches. Latter methods mostly predict heatmaps that characterize the probabilities of each keypoint at different locations [14]–[17]. The exact location of a keypoint is further estimated by finding the maximum in an aggregation of heatmaps. Compared with direct-regression methods, heatmap-based methods better leverage the distributed properties of convolutional networks and are more suitable for training human pose estimation models.

Some works combine graphical models with CNN. Tompson *et al.* [18] apply MRF as a post-processing step, while others embed deformable mixture of parts [19] or CRF [17] into the network for end-to-end learning. Convolutional Pose Machines (CPM) [15] and Stacked Hourglass Network (Hourglass) [2] achieve state-of-the-art performance without hand designed priors or graphical model-style inference. Both CPM and Hourglass employ a multi-stage scheme, using intermediate supervision to produce increasingly refined heatmaps for joints locations through different stages. The design of Stacked Hourglass Network [2] is motivated by FCN (Fully Convolutional Networks) [20] and ResNet [21]. Its powerful and well-designed architecture consists of multiple stacked hourglass modules which allow for repeated bottom-up, top-down inference [20]. Features are processed across all scales and consolidated to best capture the various spatial relationships associated with the body. Each Hourglass module contains several residual modules [21].

In order to obtain higher accuracy, image pyramids are adopted to produce multi-scale feature representations. Tompson *et al.* [18] propose a multi-branch network trained on three scales of image pyramid to learn a model with strong scale invariance. However, computation and memory will increase dramatically with the increase of scale, if image pyramids are directly used for training.

Yang *et al.* [3] propose the feature pyramid networks. They design the pyramid residual modules (PRMs) to effectively learn multi-scale deep features. The cost of computation and memory is greatly reduced.

Some works use bone-based part representations [22], [23]. The heatmaps of limbs between each pair of adjacent keypoints are taken as supervision in training.

Tang *et al.* [10] propose a Deeply Learned Compositional Model (DLCM) for Human Pose Estimation. It exploits CNNs to learn the compositionality of human bodies. The network has a hierarchical compositional architecture and bottom-up/top-down inference stages. In the bottom-up stage of DLCM, Hourglass is used to predict human pose. And subsequently the top-down stage plays the role to refine the predicted pose in bottom-up stage.

In summation: (1) Recent state-of-the-art human pose estimation methods are either improved Hourglass [3], [4], [10], or take ResNet as their backbone [5], [6], [24]; (2) Multi-scale feature can improve pose estimation; (3) Bone-based part representation can make entire body joints structural constrains more precise.

In this paper, we focus on 2D single person pose estimation from RGB images. Our work focuses on the design of bone-based part representation with feature pyramids to improve the accuracy of pose estimation.

III. OUR APPROACH

A. NETWORK ARCHITECTURE

We use the state-of-the-art hourglass architecture [2] as our base network. It is a fully convolutional network with residual modules as its building blocks. The network starts with an initial process of a 7×7 convolution with stride 2, followed by several residual modules and max-pooling layers.

The initial process reduces the resolution of the feature maps from 256 to 64. Then, a sequence of hourglass modules are stacked to predict the keypoint heatmaps.

A single hourglass module is a bottom-up and top-down design to extract the features at every scale. For human pose estimation, it is essential to explore both the local evidence, such as a small region around the wrist, and the long-range relationships between joints.

We adopt the 5-Stacked Hourglass Networks as the basic network structure. The 5-stack hourglass architecture is shown in Fig. 1. The first three Hourglass Networks implement feature pyramid learning by applying multi-scale residual modules (MSRMs), which will be described in detail in the next section (Fig.3 (b)). The Hourglass network with MSRMs is interpreted in Fig. 2. The last two hourglass networks are standard hourglass without MSRMs. At the end of each hourglass network, score maps of body joint locations are generated and a squared-error loss is computed.

In this paper, we use compositional models and bone-based part representation (See Section C, D) to represent human body joints as different levels (See Fig.4). For example, the 16 basic body joints (such as right shoulder, right elbow, etc.) belong to level-0, right upper arm (composite of right shoulder and right elbow) belongs to level-1, and right arm (composite of right upper and lower arm) belongs to level-2.

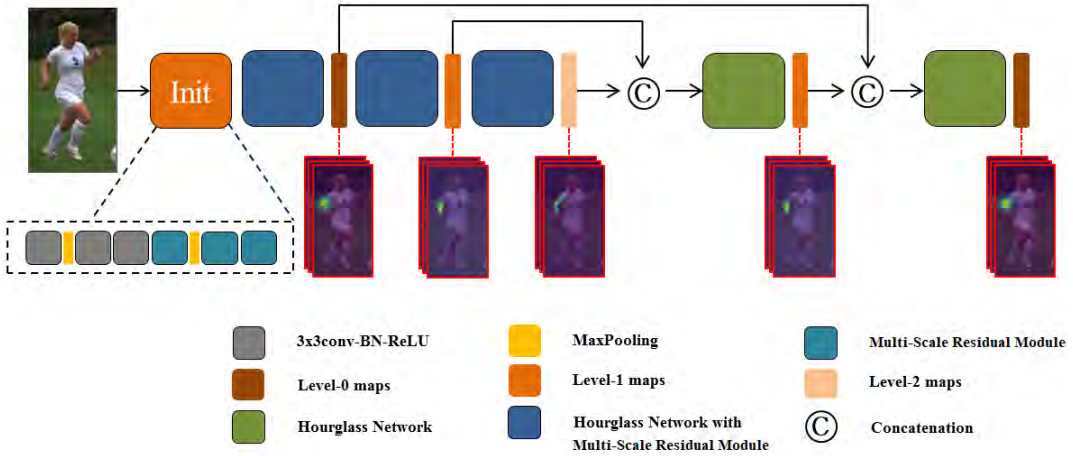

FIGURE 1. The framework of our network.

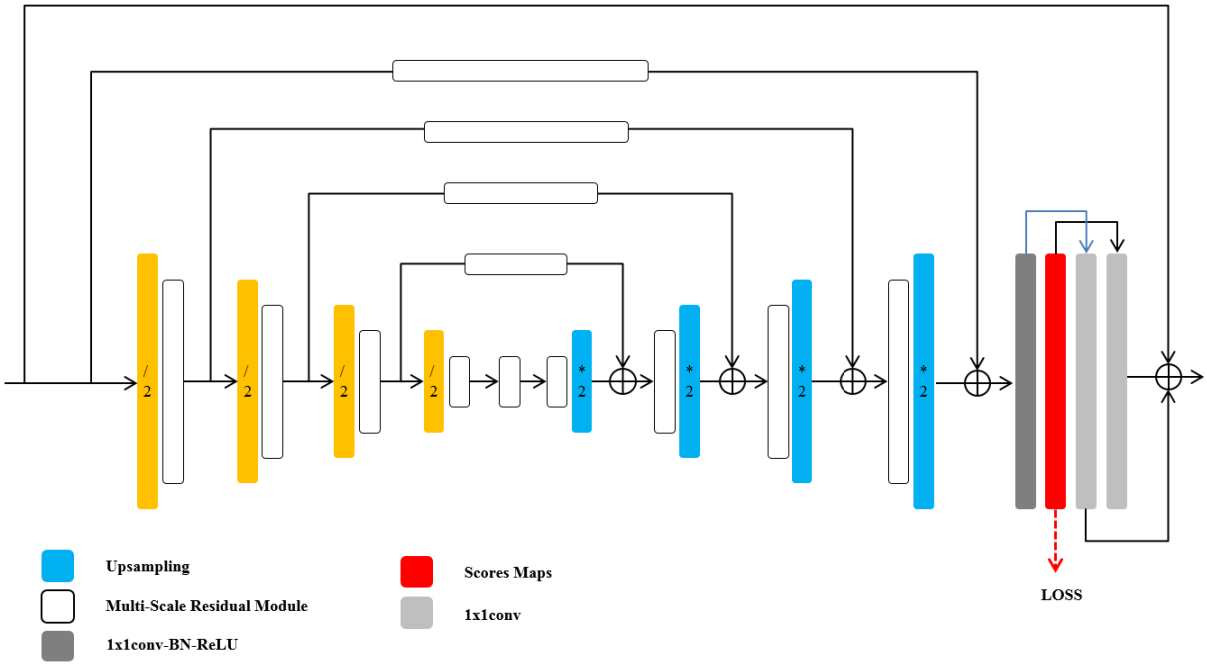

FIGURE 2. Overview of Hourglass network with multi-scale residual modules.

To maintain the information and to integrate global and local context concurrently, skip connections are used, and features at each resolution can be better preserved. In practice, the first three hourglass networks with multi-scale residual blocks play a role in bottom-up phase. Score maps of target keypoints are first regressed directly from the image observations to form level-0 maps. Then, the higher-level maps are estimated recursively from the lower-level maps. The succedent two hourglass networks enroll in a top-down phase. The lower-level maps are refined recursively using higher-level maps and high-level semantic information.

B. MULTI-SCALE RESIDUAL MODULE

For these deep convolutional neural networks based human pose estimation methods, the input images are first warped to the similar scale based on human body size. During the test phase, the test image should also be warped to the same scale as that for training images. Due to the changes of observation points and the joints of the body, the scales of different parts of the human body may still be inconsistent, which makes it difficult for the body parts detector to localize the key points.

In DCNNs, the problem of scale change occurs not only in the deeper layers with high-level semantics, but also in the shallower layers with low-level features.

In order to enhance the scale invariance, we adopt pyramid features as inputs of the network in the process of pose estimation. Specifically, the pyramid residual module (PRMs) [3] are used to construct the convolution filter of the feature pyramid. Given the input features, the pyramid residual module obtains features of different scales by sampling at different

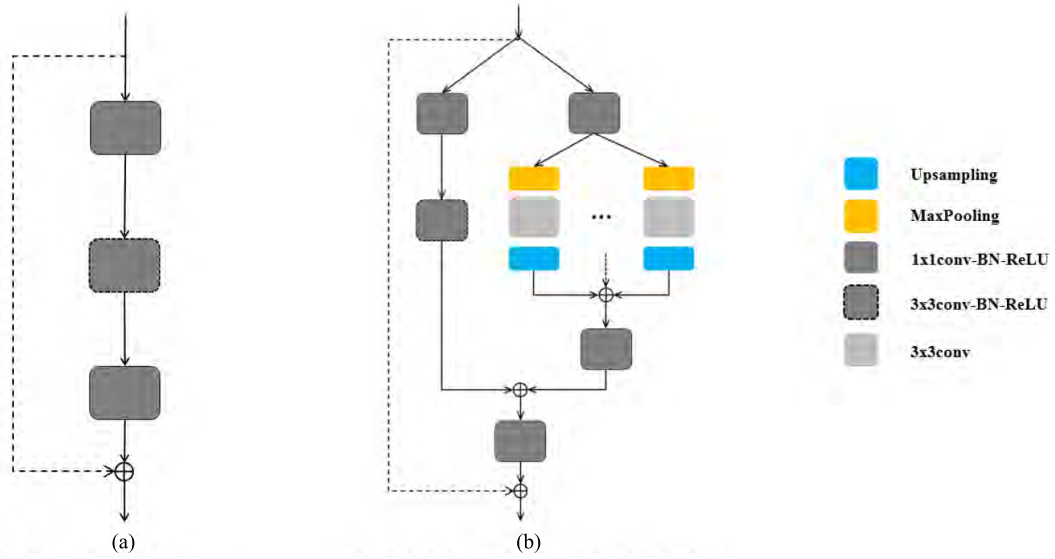

FIGURE 3. (a) Residual block is used by original hourglass to extract features. It extracts features on only one scale. (b) Multi-Scale Residual Module.

scales. Then the later convolution layer learns the features of different scales. The extracted features will be adjusted to the same resolution and summed together to form the final multi-scale feature.

The pyramid residual module can be used as a basic unit for learning different levels of features in DCNNs. Feature extraction at different levels using the pyramid residual module can produce multi-scale feature representations to achieve higher accuracy. We use the hourglass as our basic structure and replace the original residual unit with the proposed pyramid residual module, extending from a single scale to a multi-scale. We investigate how feature pyramid learning can benefit human pose estimation.

In [10], the residual units [21] are used as building blocks for the Hourglass network. However, it can only capture visual patterns or semantics on a specific scale. In this paper, we use the pyramid residual module as a building block to capture multi-scale visual patterns or semantics.

The purpose of pyramid residual module is to learn the different levels of feature pyramids, motivated by recent progress [3]. We propose a novel Deeply Learned Multi-Scale Compositional Model (DLMSCM), which is able to learn multi-scale feature pyramids of different levels of parts.

Pyramid residual module is explicitly a learning filter for input features with different resolutions. Let $x^{(l)}$ be the input of l-th layer and $W^{(l)}$ be the filter of l-th layer. The output can be written as:

$$x^{(l+1)} = x^{(l)} + M(x^{(l)}; W^{(l)}) \tag{1}$$

where $M(x^{(l)}; W^{(l)})$ is feature pyramids, can be calculated as:

$$M(x^{(l)}; W^{(l)}) = g(\sum_{i=1}^{n} f_n(x^{(l)}; w_{f_n}^{(l)}); w_g^{(l)}) + f_0(x^{(l)}; w_{f_0}^{(l)}) \tag{2}$$

where N denotes the number of branches of the feature pyramid module. $f_n(.)$ indicates the transformation of the n-th feature pyramid branch. $W^{(l)} = \{w_{f_n}^{(l)}, w_g^{(l)}\}_{n=0}^{N}$ is the branch parameters of feature pyramid module. After summing up all the output of $f_n(.)$, the final result is obtained by convolution of filter $g(.)$. An overview of the proposed framework is illustrated in Fig. 3.

To reduce computational and spatial complexity, each $f_n(.)$ is designed as a bottleneck structure. First, the feature dimension is reduced by a 1×1 convolution, and then new features are calculated on a set of sub-sampled input features using 3×3 convolutions. Finally, all new features are promoted to the same dimensions and are summed together.

C. COMPOSITIONAL MODELS

A compositional model is defined on a hierarchical graph. It consists of a 4-tuple $(\alpha, \beta, \theta^{and}, \theta^{leaf})$ that specifies its graph structure (α, β) and potential functions $(\theta^{and}, \theta^{leaf})$. α are characterized by two types of nodes: $\alpha = \alpha^{and} \bigcup \alpha^{leaf}$. α^{and} model the composition of subparts into higher-level parts and α^{leaf} represents the lowest level portion. β denotes graph edges.

For HPE, A state variable w_u can be represented by the position p_u and type t_u : $w_u = \{p_u, t_u\}, u \in \alpha$. Let φ denote the set of all state variables in the model, and it can be formulated as,

$$p(\varphi|I) = \frac{1}{T} \exp\{-E(\varphi, I)\} \tag{3}$$

where $E(\varphi, I)$ is the energy function, T is the partition function and I is the input image.

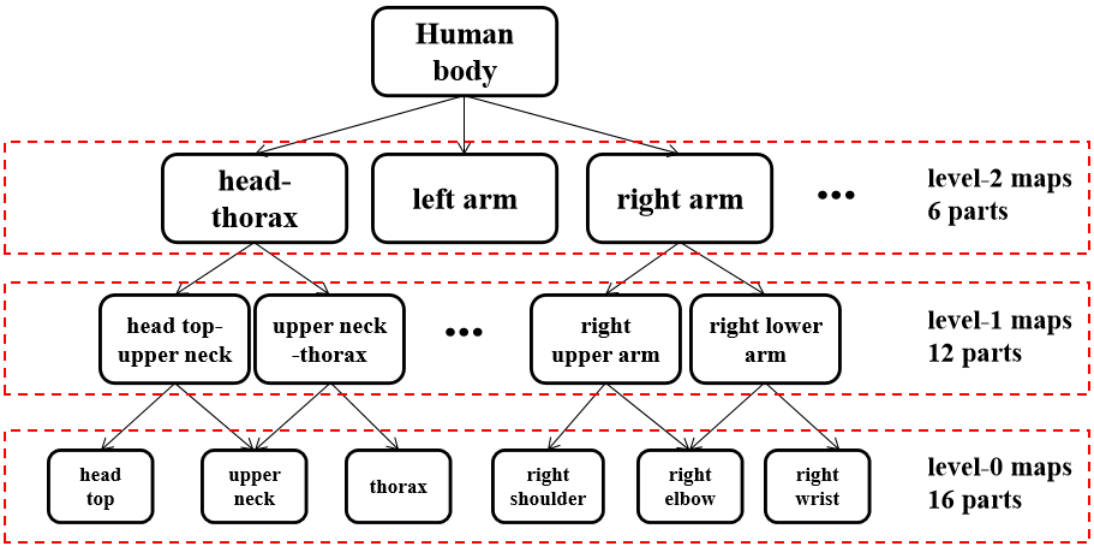

FIGURE 4. Bone-based part representation. Different bone-parts represent different levels of maps. We have three semantic levels, including 16, 12, and 6 parts.

For the simplicity of the formula, we omit I and let $H(\varphi) = -E(\varphi, I)$. $H(\varphi)$ is decomposed as:

$$H(\varphi) = \sum_{u \in \alpha^{leaf}} \theta_u^{leaf}(w_u, I) + \sum_{u \in \alpha^{leaf}} \sum_{v \in ch(u)} \theta_{u,v}^{and}(w_u, w_v) \quad (4)$$

where $ch(u)$ denotes the set of children of node u.

By using tree structure and dynamic programming, the optimal state of input image I can be calculated effectively. It is consisted of two stages: bottom-up stage and top-down stage.

In the bottom-up stage, the maximum score $H(\varphi)$ can be expressed as:

$$(Leaf)H_u^{\uparrow}(w_u) = \theta_u^{leaf}(w_u, I) \quad (5)$$

$$(And)H_u^{\uparrow}(w_u) = \sum_{v \in ch(u)} \max_{w_v} [\theta_{u,v}^{and}(w_u, w_v) + H_v^{\uparrow}(w_v)] \quad (6)$$

where $H_u^{\uparrow}(w_u)$ is the maximum score of the subgraph consisting of node u and all its subgraphs, and the state of root node u is w_u. Eq. (5) is boundary conditions.

In the top-down stage, the optimal states of child nodes can be expressed as:

$$(Root)w_u^* = argmax_{w_u} H_u^{\downarrow}(w_u) \equiv argmax_{w_u} H_u^{\uparrow}(w_u) \quad (7)$$

$$(Non-root)w_v^* = argmax_{w_v} H_v^{\downarrow}(w_v) \equiv argmax_{w_v} [\theta_{u,v}^{and}(w_u^*, w_v) + H_v^{\uparrow}(w_v)] \quad (8)$$

The node u in Eq. (8) is the only parent node of node v, $H_u^{\uparrow}(w_u)$ and $H_v^{\uparrow}(w_v)$ are obtained from bottom-up stage, and $H_u^{\downarrow}(w_u)$ and $H_v^{\downarrow}(w_v)$ are the refinement score graphs of nodes u and v, respectively. Especially w_u^* and w_v^* are respectively the optimal state of u and v. Eq. (7) is boundary conditions.

D. BONE-BASED PART REPRESENTATION

As shown in Fig. 4, we use the bones to represent each part, which is generated by placing a Gaussian kernel along a key point. Then, when training the model, the score graphs $H_u^{\uparrow}(w_u)$ and $H_u^{\downarrow}(w_u)$ are taken as the ground truth of these parts. Specifically, we generate a heat map centered on 2D Gaussian (std = 1 pixel) at each point on the line segment of the part.

E. DEEPLY LEARNED MULTI-SCALE COMPOSITIONAL MODEL (DLMSCM)

Our Deeply Learned Multi-Scale Compositional Model (DLMSCM) uses feature pyramid and deep compositional models to learn multi-scale features and full body joints constraints. In the bottom-up stage, the score maps of the key points are first regressed from the input image. Then, the score maps of those higher levels are estimated recursively from their child nodes. In this stage, feature pyramids are used to acquire multi-scale features to make regression more accurate. In the top-down stage, the score maps of lower-level parts are refined recursively by using their parents' score maps and their own scoring maps estimated in the bottom-up stage.

IV. EXPERIMENTS

A. DATASETS AND IMPLEMENTATION DETAILS

Our approach is evaluated on two widely used human pose estimation benchmarks: Leeds Sports Pose (LSP) [25] and MPII Human Pose Dataset [26]. The LSP and its extended dataset [25] consist of 11k training images and 1k testing images. Each image is annotated with 14 keypoints, which is gathered from Flickr. One of the challenges of the dataset is noisy labels. Another challenge of the dataset is variety of

poses from sports activities such as baseball, parkour, tennis, and so on. MPII dataset [26] contains about 25k images with over 40k annotated people, which covers a great variety of human activities. Each image is annotated with 16 joints, the center and scale. As previous works [3], we train the network by including the MPII training samples. In comparison with other pose datasets, MPII dataset are more complex in terms of human interaction and poses.

Each input image is cropped 256×256 from resized images according to the target annotated human body center and scal. Each training image is augmented by scaling, rotation, flipping, and adding color noise. Our models use Torch7 deep learning libraries. We use RMSProp to optimize the network with a mini-batch size of 16 for 250 epochs. Training is performed on two 16GB NVIDIA Tesla P100 GPUs. The learning rate is initialized as 1×10^{-4} and dropped by 10 at 180th and 225th epoch.

B. EVALUATION METRICS

Following previous works [2]–[4], we use Percentage of Correct Keypoints (PCK) [1] to evaluate performance on LSP dataset, and use PCKh [26] on the MPII dataset.

1) PERCENTAGE OF CORRECT KEYPOINTS (PCK)

PCK reports the percentage of correct detections that fall within a normalized distance. The distance is calculates by the torso size. Let $\hat{z}_k$ be the predicted location of the kth body joint, z_k is the corresponding ground truth location, then PCK is defined as:

$$\frac{||z_k - \hat{z}_k||_2}{||z_{lhip} - z_{rsho}||_2} \leq r \tag{9}$$

where z_{lhip} and z_{rsho} denote the ground truth locations of the left hip and right shoulder, respectively. $r \in [0, 1]$ is a threshold.

2) PERCENTAGE OF CORRECT KEYPOINTS WITH RESPECT TO HEAD (PCKH)

Similar to PCK, The distance is normalized by a fraction of head size, and the measure is referred to as PCKh.

C. RESULTS

1) QUANTITATIVE RESULTS

In our experiments, we evaluate our method on two standard benchmarks MPII and LSP. We compare our approach with other state-of-the-art methods, the results are listed in Table 1 and Table 2.

Table1 shows the comparisons of the PCK scores at the threshold of 0.2 (PCK@0.2) on LSP test set. As previous works [3], [10], our models are trained by adding MPII training set to LSP training and LSP extended training set. As for the mean accuracy of key points, our approach achieves the new best performance of 95.3%, and improves the previous best result [10] by 0.2%. Our method achieves the best scores on six body parts, head, Elbow, wrist, Hip, Knee and Ankle with 98.6%, 93.8%, 90.8%, 95.3%, 97.0% and 95.8% respectively.

TABLE 1. Comparisons of PCK@0.2 scores on the LSP testing set.

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Wei et al. [15], CVPR'16	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5
Bulat et al. [27], ECCV'16	97.2	92.1	88.1	85.2	92.2	91.4	88.7	90.7
Chu et al. [17], CVPR'17.	98.1	93.7	89.3	86.9	93.4	94.0	92.5	92.6
Yang et al.[3], ICCV'17	98.3	94.5	92.2	88.9	94.7	95.0	93.7	93.9
Ning et al.[28], TMM'17	98.2	94.4	91.8	89.3	94.7	95.0	93.5	93.9
Chou et al.[4], arXiv'17	98.2	94.9	92.2	89.5	94.2	95.0	94.1	94.0
Tang et al.[10], ECCV'18	98.3	95.9	93.5	90.7	95.0	96.6	95.7	95.1
Ours	98.6	95.8	93.8	90.8	95.3	97.0	95.8	95.3

* The red numbers indicate the best scores, and the blue ones are second.

TABLE 2. Comparisons of PCKh@0.5 scores on the MPII testing set.

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Bulat et al. [27], ECCV'16	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Newell et al. [2], ECCV'16	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Tang et al. [29], ECCV'18	97.4	96.4	92.1	87.7	90.2	87.7	84.3	91.2
Ning et al. [28], TMM'17	98.1	96.3	92.2	87.8	90.6	87.6	82.7	91.2
Chu et al. [17], CVPR'17	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
Chou et al.[4], arXiv'17	98.2	96.8	92.2	88.0	91.3	89.1	84.9	91.8
Chen et al.[9], ICCV'17	98.1	96.5	92.5	88.5	90.2	89.6	86.0	91.9
Yang et al.[3], ICCV'17	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0
Ke et al.[30], ECCV'18	98.5	96.8	92.7	88.4	90.6	89.3	86.3	92.1
Tang et al.[10], ECCV'18	98.4	96.9	92.6	88.7	91.8	89.4	86.2	92.3
Ours	98.5	97.1	92.7	88.9	91.6	89.6	86.6	92.5

* The red numbers indicate the best scores, and the blue ones are second.

In Table 2, we give the comparisons of the PCKh scores at the threshold of 0.5(PCKh@0.5) on MPII test set. Our approach achieves 92.5%, which clearly outperforms the previous work and improves the previous best result [10] by 0.2%. In addition, Our approach achieves the best scores on six body parts, head, Shoulder, Elbow, wrist, Knee and Ankle. Compared with the counterpart method [10] whose total PCKh@0.5 is 92.3%, our method acquires 0.2% improvement by taking advantage of feature pyramids. Specifically, our methods achieve 0.2% improvements on shoulder, wrist and knee. For the most challenging parts to be detected as Ankle, our improvements are even notable, with 0.3%.

Complexities: Table.3 compares the complexity of our DLMSCM model with other current state-of-the-art approaches. Obviously, the parameters of DLMSCM model

FIGURE 5. Qualitative comparisons on LSP.

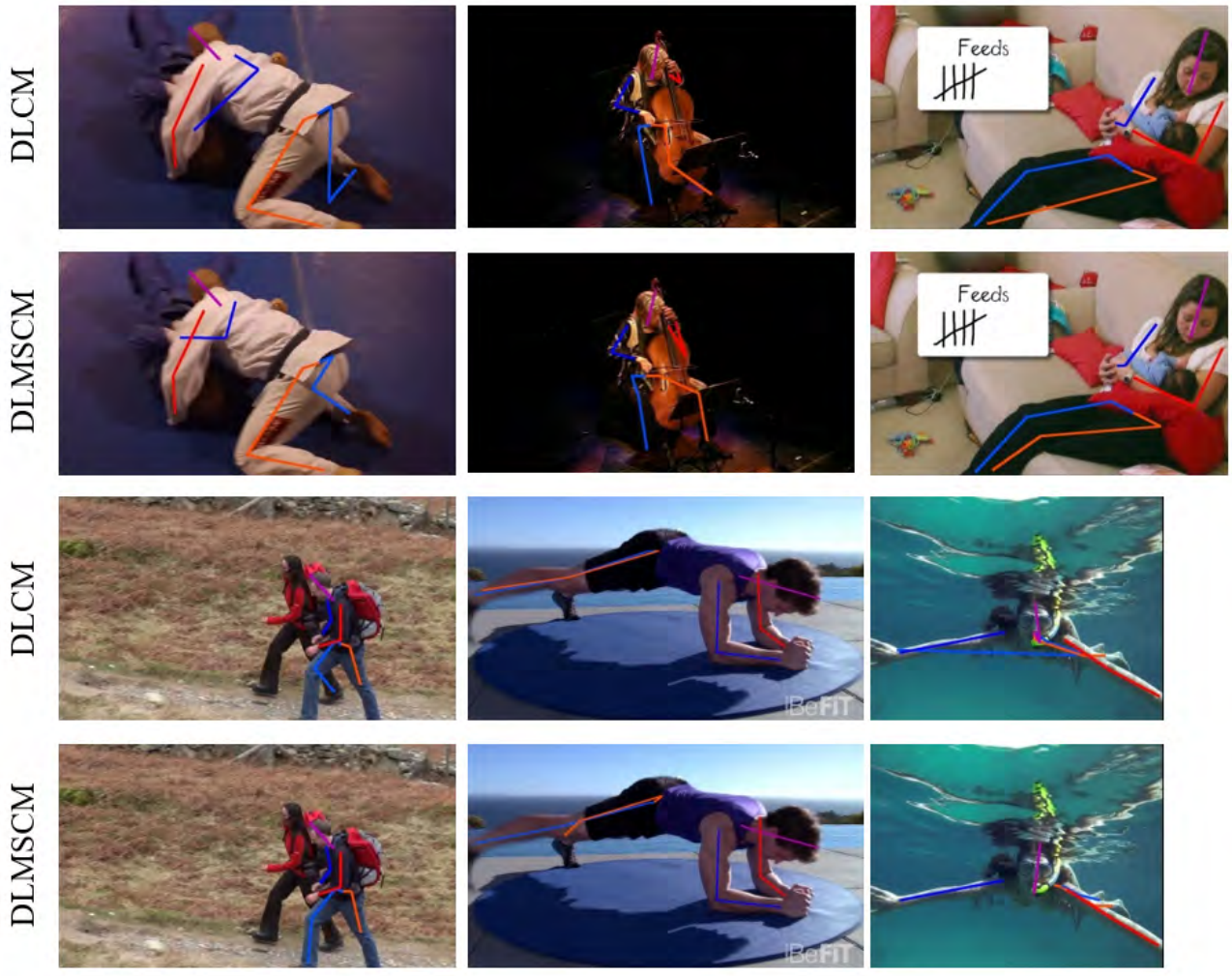

FIGURE 6. Qualitative comparisons on MPII.

is similar to that of Tang *et al.* [10]. However, our method has fewer parameters and lower computational complexity than the methods of Yang *et al.* [3] and Newell *et al.* [2].

2) QUALITATIVE COMPARISONS

We show some qualitative comparison results on LSP in Fig. 5, and on MPII in Fig. 6. We compare our approach

TABLE 3. Comparisons of parameters and model size.

	Parameters	Model size
Yang et. al. [3]	26.9M	191M
Newell et. al. [2]	23.7M	196.08M
Tang et. al. [10]	15.5M	119.25M
Ours	18.34M	133.16M

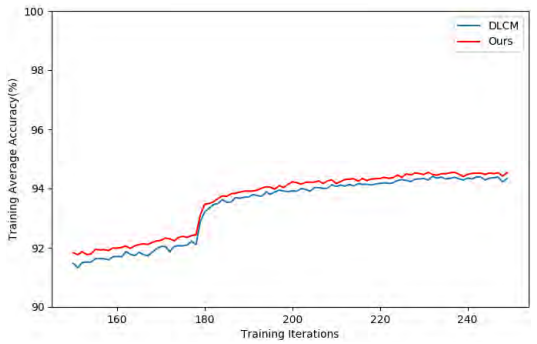

FIGURE 7. Training curves of PCK scores vs. epoch on the LSP training set. The blue curve stands for the DLCM model, and the red curve represents our model.

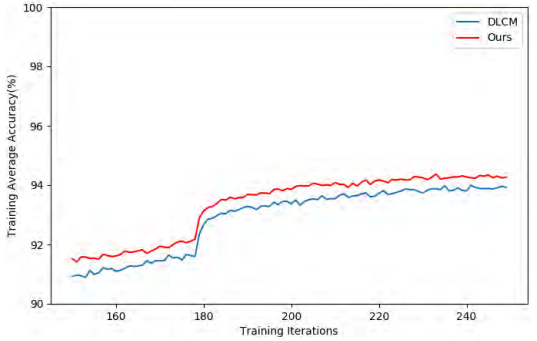

FIGURE 8. Training curves of PCKh scores vs. epoch on the MPII training set. The blue curve stands for the DLCM model, and the red curve represents our model.

with the DLCM approach. In each figure, the first row contains some results predicted by DLCM [10], and our results are in the second row.

We can see that, by utilizing multi-scale mechanism in the framework of DLCM, the model can be refined to correct some errors to produce more plausible poses [3]. For example, from the first to the fourth columns in Fig. 5, the DLCM model can not distinguish heavily confusing joints, while our model can learn the correct joint relationships well. Further, multi-scale mechanism can enhance structural dependencies among body joints. For example, in the last column of Fig. 6, the DLCM model fails in case of complex occlusions, while our model can successfully predict the correct joint locations. The performances are obviously improved.

To further analyze the performance of each approach with respect to the same factors, we show the training accuracy curve in Fig. 7 and Fig. 8. We zoom in the part of curve after epoch 150. We find that the strategy of learning rate decay is helpful for both methods. At the 180th epoch, with the decrease of learning rate, the PCKh scores on the MPII training set and the PCK scores on the LSP training set increase significantly. However, there are some differences. Ours is a bit more stable and always achieves better performance.

V. CONCLUSION

This paper proposes to learn feature pyramids by multi-scale residual modules in Deeply Learned Compositional Model. Considering the abundant dependency information among body joints, we use bone-based part representation to combine and describe the relevant joints. In order to detect and locate the skeleton and parts more accurately, a training method based on different level-maps is proposed. The level-maps are generated by Gaussian kernel to represent bone-based body parts. We use multi-scale residual modules to enhance the invariance of scales of the complex compositional models for human pose estimation. Experimental results on MPII and LSP dataset demonstrate the effectiveness of our proposed method.

REFERENCES

[1] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2878–2890, Dec. 2013.
[2] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. ECCV*, 2016, pp. 483–499.
[3] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *Proc. ICCV*, Oct. 2017, pp. 1290–1299.
[4] C.-J. Chou, J.-T. Chien, and H.-T. Chen, "Self adversarial training for human pose estimation," 2017, *arXiv:1707.02439*. [Online]. Available: https://arxiv.org/abs/1707.02439
[5] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proc. CVPR*, Jun. 2018, pp. 7103–7112.
[6] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. ECCV*, 2018, pp. 466–481.
[7] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *Proc. CVPR*, Jun. 2016, pp. 4733–4742.
[8] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proc. CVPR*, Jul. 2017, pp. 2261–2269.
[9] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang, "Adversarial posenet: A structure-aware convolutional network for human pose estimation," in *Proc. ICCV*, Oct. 2017, pp. 1221–1230.
[10] W. Tang, P. Yu, and Y. Wu, "Deeply learned compositional models for human pose estimation," in *Proc. 15th ECCV*, Munich, Germany, Sep. 2018, pp. 190–206.
[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
[13] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. CVPR*, Jun. 2014, pp. 1653–1660.
[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.
[15] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. CVPR*, Jun. 2016, pp. 4724–4732.
[16] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. CVPR*, Jul. 2017, pp. 1302–1310.

[17] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. CVPR*, Jul. 2017, pp. 5669–5678.
[18] J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Proc. NIPS*, 2014, pp. 1799–1807.
[19] W. Yang, W. Ouyang, H. Li, and X. Wang, "End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation," in *Proc. CVPR*, Jun. 2016, pp. 3073–3082.
[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, Jun. 2015, pp. 3431–3440.
[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Jun. 2016, pp. 770–778.
[22] B. Ai, Y. Zhou, Y. Yu, and S. Du, "Human pose estimation using deep structure guided learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2017, pp. 1224–1231.
[23] V. Belagiannis and A. Zisserman, "Recurrent human pose estimation," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, May/Jun. 2017, pp. 468–475.
[24] X. Liang, K. Gong, X. Shen, and L. Lin, "Look into person: Joint body parsing & pose estimation network and a new benchmark," Apr. 2018, *arXiv:1804.01984*. [Online]. Available: https://arxiv.org/abs/1804.01984
[25] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *Proc. BMVC*, 2010, p. 5.
[26] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. CVPR*, Jun. 2014, pp. 3686–3693.
[27] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *Proc. ECCV*, 2016, pp. 717–732.
[28] G. Ning, Z. Zhang, and Z. He, "Knowledge-guided deep fractal neural networks for human pose estimation," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1246–1259, May 2018.
[29] Z. Tang, X. Peng, S. Geng, L. Wu, S. Zhang, and D. Metaxas, "Quantized densely connected U-nets for efficient landmark localization," in *Proc. ECCV*, 2018, pp. 339–354.
[30] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, "Multi-scale structure-aware network for human pose estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 731–746.