Pattern Recognition 106 (2020) 107410

Contents lists available at ScienceDirect

Pattern Recognition

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

Structure-aware human pose estimation with graph convolutional networks

Yanrui Bin[a], Zhao-Min Chen[b], Xiu-Shen Wei[c,*], Xinya Chen[a], Changxin Gao[a], Nong Sang[a,*]

[a] Key Laboratory of Image Processing and Intelligent Control, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China
[b] National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[c] Megvii Research Nanjing, Megvii Technology Ltd., Nanjing, China

ARTICLE INFO

Article history:
Received 28 March 2019
Revised 19 December 2019
Accepted 29 April 2020
Available online 16 May 2020

Keywords:
Human pose estimation
Graph convolutional networks
Key points structural relations

ABSTRACT

Human pose estimation is the task of localizing body key points from still images. As body key points are inter-connected, it is desirable to model the structural relationships between body key points to further improve the localization performance. In this paper, based on original graph convolutional networks, we propose a novel model, termed Pose Graph Convolutional Network (PGCN), to exploit these important relationships for pose estimation. Specifically, our model builds a directed graph between body key points according to the natural compositional model of a human body. Each node (key point) is represented by a 3-D tensor consisting of multiple feature maps, initially generated by our backbone network, to retain accurate spatial information. Furthermore, attention mechanism is presented to focus on crucial edges (structured information) between key points. PGCN is then learned to map the graph into a set of structure-aware key point representations which encode both structure of human body and appearance information of specific key points. Additionally, we propose two modules for PGCN, i.e., the Local PGCN (L-PGCN) module and Non-Local PGCN (NL-PGCN) module. The former utilizes spatial attention to capture the correlations between the local areas of adjacent key points to refine the location of key points. While the latter captures long-range relationships via non-local operation to associate the challenging key points. By equipping with these two modules, our PGCN can further improve localization performance. Experiments both on single- and multi-person estimation benchmark datasets show that our method consistently outperforms competing state-of-the-art methods.

1. Introduction

Human pose estimation is the task of localizing body key points from still images. It serves as a fundamental technique for numerous computer vision applications, such as action recognition [1–4], person re-identification [5], human-computer interaction and so on. It is also a challenging problem due to the high flexibility of body limbs, occlusions, clutter backgrounds, overlapping parts, nearby persons etc.

A Naïve way to address the pose estimation problem is to treat the body key points in isolation and predict a set of heat maps which produce a per-pixel likelihood for key points locations using powerful convolutional neural networks (CNNs). Recently, Wei et al. [6], Newell et al. [7] improves isolated detection based methods by multi-stage prediction mechanism, where predictions produced by the previous stage are fed to the next stage to learn the image-dependent spatial distribution of key points. However, as shown in Fig. 1, these methods are prone to fail in the challenging cases. One promising way to further improve the localization performance is to exploit the structural relationships between key points. Some works based on probabilistic graph model [8,9] are proposed to learn typical spatial relationships between key points. However the feature correlation between key points is ignored.

In this paper, based on powerful graph convolutional networks, we propose a Pose Graph Convolutional Network (PGCN) to capture the structural relationships between key points for pose estimation. Specifically, our PGCN represents the node (key points) feature description by a 3-D tensor consisting of multiple 2-D feature maps to retain accurate localization information, and builds a directed graph over these key points representations to explicitly model their correlations later. Each layer of PGCN first transforms the input key points feature maps into a higher-level feature space by convolutions, and then employs attention mechanisms to focus on crucial edges (structured information) between key points

* Corresponding authors.
E-mail addresses: weixs.gm@gmail.com (X.-S. Wei), nsang@hust.edu.cn (N. Sang).

https://doi.org/10.1016/j.patcog.2020.107410
0031-3203/

Fig. 1. Pairs of pose predictions obtained by an 8-stack Hourglass network [7] (left) and our PGCN model (right). By capturing and exploiting structure of human body, our model generates more accurate results.

which vary with input poses, type of key points and spatial locations. A multi-layer PGCN is stacked to generate a set of structure-aware key point representations which encode both structure of human body and appearance information of key points. These key point representations are used to predict key point heat maps which indicate the localization of each key point. Furthermore, we develop two types of PGCN modules, namely the Local PGCN (L-PGCN) module and Non-Local PGCN (NL-PGCN) module. Local PGCN uses the spatial attention to pass the messages between the local areas of adjacent key points. The detailed description generated by L-PGCN is beneficial for accurate localization of the body key points. Non-Local PGCN employs non-local operations to model the relationships in spite of the position of key points, which enables the network to effectively handle challenging key points. Ultimate aggregation of the heat maps generated by both L-PGCN and NL-PGCN achieves superior performance.

The main contributions of this paper are as follows:

- We propose a novel pose estimation model, which leverages the power of graph convolutional networks to explicitly model the structured relationships between key points.
- We design two modules for our model, i.e., Local PGCN and Non-Local PGCN which are proposed to refine the location of key points locally, and capture global underlying contextual information, respectively.
- We comprehensively evaluate our model on two single-person pose estimation datasets and a multi-person estimation dataset, and our proposed method consistently achieves superior performance over previous state-of-the-art approaches.

The remainder of the paper is organized as follows. In Section 2, we briefly review related literature of pose estimation and graph convolutional networks. In Section 3, we elaborately introduce our proposed PGCN model. Section 4 reports the pose estimation results of both single- and multi-person benchmark datasets with also ablation studies. We conclude the paper in Section 5 finally.

2. Related work

Our proposed approach is related to previous work on graph convolutional networks and human pose estimation, which is categorized as single- and multi-person pose estimation.

2.1. Single-person pose estimation

Single-person pose estimation has been an active research area. Early approaches modeled human body as a set of unary term and pairwise term. The unary term captured part appearance using hand-craft feature such as histogram of oriented gradients (HOG) while pairwise term captured spatial relationships among parts. The majority of early work [10–12] focused on proposing an strong pairwise term for highly articulated human body.

Recently, pose estimation using CNNs has shown superior performance, which can be categorized into two categories: regression based and detection based. Regression based methods [13] directly regressed the 2D coordinates of key points from the input image. Nevertheless, they are not performing as well as detection based methods due to its lack of inherent spatial generalization.

Detection based methods predicted a heat map for each key point and located the key point as the point with the maximum value in the map. Early works [8,9,14] focused on exploiting structural constraints between key point locations to solve the multi-mode problem of heat map representation. Tompson et al. [8] jointly trained a MRF-based spatial model and a multi-resolution CNNs. The method of [14] proposed geometrical transform kernels to capture the relationships between feature maps of key points. Yang et al. [9] combined CNNs with the expressive deformable mixture of parts. CPM [6] used a sequential composition of convolutional architectures with large receptive field to learning an implicit spatial models. Newell et al. [7] followed CPM's framework and designed stacked hourglass network to rapidly expand the receptive field and consolidate features from various scales to best capture the various spatial relationships associated with the body.

Since then occlusion and ambiguity became the main difficulty of single-person pose estimation and various methods [15,16], based on hourglass network, were proposed to further model and exploit the relations between key points. Chen et al. [16] trained the generator (pose estimation network) in an adversarial manner against the discriminator. Tang et al. [15] introduced Deeply Learned Compositional Model (DLCM) to learn the compositionality of human bodies. However, more should be done to model and exploit the priors of the human body structure in pose estimation.

2.2. Multi-person pose estimation

Multi-person pose estimation, which involves simultaneously detecting people and localizing their key points, has been attracting intensive interests in both academia and industry. Advances in this topic are often categorized into bottom-up and top-down approaches.

2.2.1. Bottom-up fashion

Bottom-up approaches directly detect key points first and assign them to person instances. State-of-art methods use CNN to predict body parts and group assignments simultaneously, and then employ an assignment algorithm to form individual skeletons. DeepCut [17] used Integer Linear Program (ILP) to select and label body part candidates, and partitioned them into person clusters. Cao et al. [18] presented Part Affinity Fields (PAFs) and used Hungarian bipartite matching algorithm to efficiently associate key points with individuals in the image. Li et al. [19] improves PAFs [18] by parsing the poses with bounding box constraints in a top-down manner. Kocabas et al. [20] used a multi-task model to simultaneously produce score maps and person detection results, and then used a Pose Residual Network (PRN) to group the candidate key points to different people. PoseAE [21] output an associative embedding to identify key points from the same person. Zhao et al. [22] improved PoseAE by predicting tag embedding cluster-wise.

2.2.2. Top-down fashion

Top-down approaches interpret the process of detecting key points as sequentially performing person detection and single-person pose estimation. Papandreou et al. [23] used ResNet with dilated convolutions and predicted both key points heat map and offset output, which were aggregated by Hough voting to produce highly localized activation maps. MASK RCNN [24] firstly predicted person box and appended a key point head on ROI aligned feature maps to generate a one-hot mask for each key points. RMPE [25] proposed a symmetric spatial transformer network (SSTN) to handle inaccurate bounding box and introduced a parametric NMS to delete redundant detection. CPN [26] proposed by Chen et al. combined GlobalNet, which was a feature pyramid network to handle easy key points, with RefineNet to explicitly handling the hard key points by integrating all levels of feature representations from the GlobalNet together with an online hard key point mining loss. Xiao et al. [27] proposed a simple model which simply added a few de-convolution layers over the last convolution stage in the ResNet and achieve state-of-art performance.

2.3. Graph convolutional network

There is an increasing interest in generalizing convolutions to the graph domain. Advances in this direction are often categorized as spectral approaches and non-spectral approaches. Spectral approaches [28] work with a spectral representation of the graphs. The convolution operation was defined in the Fourier domain by computing the eigendecomposition of the graph Laplacian. Non-spectral approaches defined convolutions directly on the graph, operating on spatially close neighbors. Hamilton et al. [29] proposed the GraphSAGE which generated embeddings by sampling and aggregating features from local neighborhood nodes. Recently, graph convolutional networks were explored in a wide range of area such as image classification [30], text classification [31], traffic forecasting [32] and emotion distribution learning [33]. Specifically, Chen et al. [30] built a directed graph where each node corresponds to an object labels and took the word embeddings of nodes as input for predicting the classifier of different categories. Yao et al. [31] regarded the documents and words as nodes and used the Text GCN to learning embeddings of words and documents. Zhang et al. [32] represented roads and intersctions as nodes and took the summation of the taxi flow in previous six internals of nodes as input to predict the taxi flow in next internals. He and Jin [33] built a directed graph between different emotions to capture the co-appearance correlation.

3. Proposed method

An overview of the proposed framework is illustrated in Fig. 2. In this section, we recap the original graph convolutional networks and then elaborate our PGCN for structure-aware pose estimation. Furthermore we describe the two major components in our model, i.e., Local PGCN module and Non-Local PGCN module. The former focuses on the local areas of adjacent nodes feature maps to refine the location of key points, while later captures the feature correlation in spite of the position of key points which enables the network to efficiently associate challenging key points.

3.1. Graph convolution network recap

Graph Convolutional Networks (GCNs) were introduced in [34] to perform semi-supervised classification on graph-structured data. The essential idea is to update the node representations by propagating information between nodes.

Different from standard convolutional operations, the goal of GCNs is to learn a function $f(\cdot)$ on a graph $\mathcal{G}$ which takes an adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and a feature description $\boldsymbol{z}_u^l \in \mathbb{R}^d$ for every node u at lth layer as inputs. Let $\boldsymbol{z}^l$ denote the $n \times d$ feature matrix obtained by stacking together all the node feature description of the graph $\mathcal{G}$. n is the number of nodes and d is the dimension of features. Then it produces a node-level output $\boldsymbol{z}_u^{l+1} \in \mathbb{R}^{d'}$ for every node u. Every neural network layer can then be written as a nonlinear function:

$$\boldsymbol{z}^{l+1} = f^l(\boldsymbol{z}^l, \boldsymbol{A})\,. \tag{1}$$

Specifically, [34] introduced a sophisticated form of a layer-wise propagation rule which update the set of features $\boldsymbol{z}_u^l$ by two steps. Firstly a linear transformation $\mathcal{T}^l(\cdot)$, parametrized by a weight matrix $\boldsymbol{\Theta}^l \in \mathbb{R}^{d'\times d}$, is applied to transform current feature into a higher-level features $\boldsymbol{b}_u^l \in \mathbb{R}^{d'}$ and then the normalized adjacency matrix $\hat{\boldsymbol{A}}$ is used to aggregate information from its neighborhood $\boldsymbol{N}_u$. Formally, node features is updated as:

$$\boldsymbol{b}_u^l = \mathcal{T}^l(\boldsymbol{z}_u^l; \boldsymbol{\Theta}^l) = \boldsymbol{\Theta}^l \boldsymbol{z}_u^l\,, \tag{2}$$

$$\boldsymbol{z}_u^{l+1} = \sigma\left(\sum_{v\in\mathcal{G}} \hat{\boldsymbol{A}}_{u,v} \boldsymbol{b}_v^l\right), \tag{3}$$

where $\sigma(\cdot)$ denotes a nonlinear function such as ReLU [35]. Thus, we can learn and model the complex inter-relationships of the nodes by stacking multiple graph convolution layers.

3.2. Local PGCN

In this section, we design a new GCN namely PGCN for pose estimation. Intuitively, each node in $\mathcal{G}$ represents a key point and each edge connects two adjacent key points. Fig. 5 illustrates the natural compositional model of a human body.

Since human pose estimation requires accurate localization of body key points, the input key point feature description of our PGCN is a set of feature maps $\mathbf{Z}_u^0 \in \mathbb{R}^{H\times W\times C}$, which are generated by backbone network, to retain accurate spatial information. H, W and C denote the height, width and number of channels of $\mathbf{Z}_u^0$ respectively. For the output, we predict a new set of structure-aware key point representations $\mathbf{Z}_u^L \in \mathbb{R}^{H\times W\times C}$ which encode both appearance information of specific key points and structure of human body. L is the number of PGCN layer. $L = 2$ is used by default unless otherwise noted.

Consequently, function $\mathcal{T}^l$ applied to each node is naturally implemented by a convolutional operation. Note that transformation function on each node is not shared. The reason is that different convolutions for different nodes allow the PGCN to capture more accurate structural information from various human pose when a large amount of training data with the same graph structure are available. Formally, transformation $\mathcal{T}_u^l$ for each node u can be rewritten:

$$\mathbf{B}_u^l = \mathcal{T}_u^l\left(\mathbf{Z}_u^l; \boldsymbol{\Theta}_u^l\right) = \boldsymbol{\Theta}_u^l * \mathbf{Z}_u^l\,, \tag{4}$$

where "$*$" denotes convolution, $\mathbf{B}_u^l \in \mathbb{R}^{H\times W\times C'}$ is the output of $\mathcal{T}_u^l$ and $\boldsymbol{\Theta}_u^l$ is the convolution weight of node u. In all the experiments, we set $C = C' = 16$.

After transformation, directly aggregating information with the normalized adjacency matrix $\hat{\boldsymbol{A}}$ will cause several problem [50]. In the one hand, relationships between key points vary as the input pose changes. For example, relationship between a visible key point and a occluded key point is different from the one between two visible key points. In the other hand, different key points need different information from its neighborhoods. Some easy key points like eyes should be influenced less by its neighborhood than the hard one such as wrists and ankles. Furthermore, relationships of nodes features at different positions are also varied.

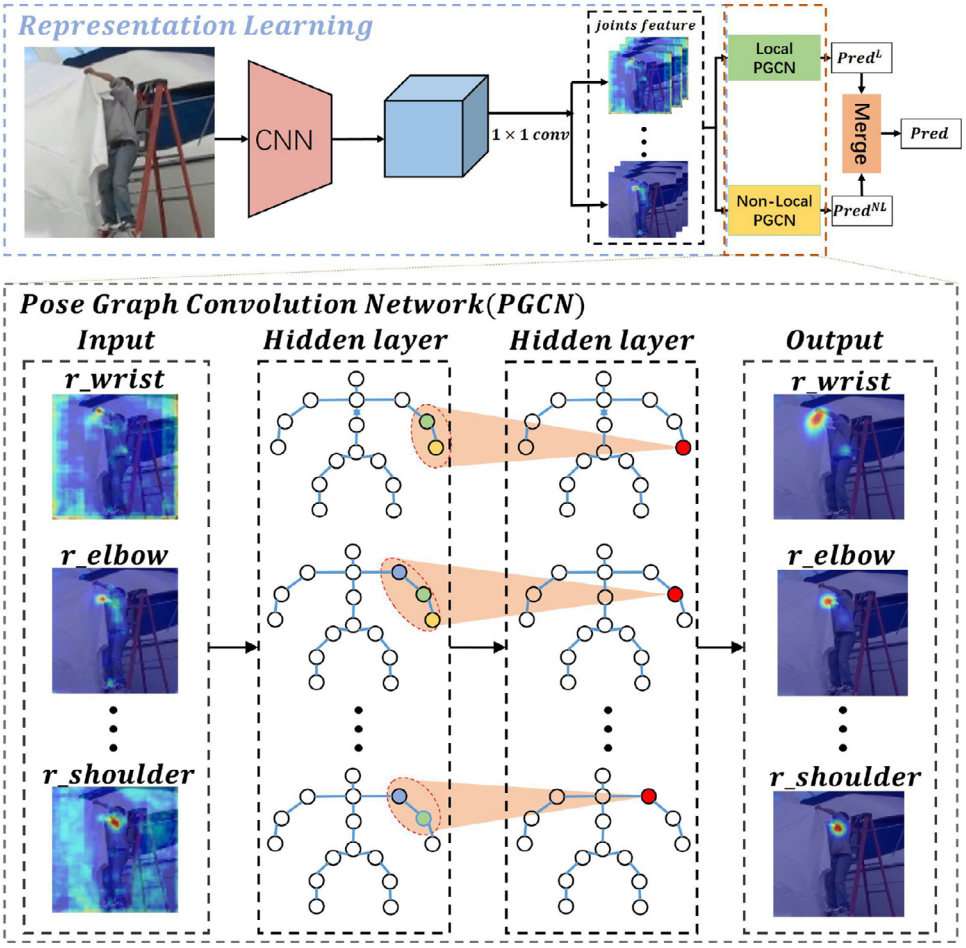

Fig. 2. Overall framework of our proposed model for human pose estimation. The input images are fed to conventional CNNs to obtain the key points representations each of which is a feature map $\mathbf{Z}_u^0 \in \mathbb{R}^{H\times W\times C}$. H, W and C are the height, width and number of channels of $\mathbf{Z}_u^0$ respectively. Then we build a directed graph over these key point representations to explicitly model their relationships. Based on the graph, two parallel multi-layer Pose Graph Convolutional Network (PGCN) modules are learned to propagate information between different key points and further exploit the key point dependency. Specifically, Local PGCN (L-PGCN) module captures the feature correlation between key points by focusing on local receptive field to refine the location of key points, while Non-Local PGCN (NL-PGCN) module exploits the long-range relationships via non-local operation to associate challenging key points. In consequence, both PGCN modules generate refined key point representations which encode both structure of human body and appearance information of key points. Two sets of heat maps are generated via applying 3×3 convolutions on the feature maps and then merged together to get our final predictions.

For above considerations, we design a attention based aggregation function to focus on crucial edges (structured information) between nodes (key points). Specifically, we perform a convolution with ReLU nonlinearity to generate an attention map $\mathbf{S}_{u,v} \in \mathbb{R}^{H\times W\times 1}$:

$$\mathbf{S}_{u,v}^l = \sigma\left(\boldsymbol{att}_{u,v}^l * concate(\mathbf{B}_u^l, \mathbf{B}_v^l)\right), \tag{5}$$

where $\boldsymbol{att}_{u,v}^l$ denotes the convolution filters and $\boldsymbol{N}_u$ denotes the adjacent key point set of key point u. $\mathbf{S}_{u,v}^l$ indicates the importance of node v to node u at every positions. Here, $\boldsymbol{att}_{u,v}^l$ is specific for each (u, v). The layer-wise propagation rule is illustrated Fig. 3(a).

Once obtained, attention map is used to generate a linear combination of the feature map corresponding to them, to serve as the output for each node:

$$\mathbf{Z}_u^{l+1} = \sum_{v\in\boldsymbol{N}_u} \mathbf{S}_{u,v}^l \odot \mathbf{B}_v^l, \tag{6}$$

where "$\odot$" represents the channel-wise Hadamard matrix product operation. Note that $\boldsymbol{N}_u$ contains node u itself. Specifically, K independent attention mechanisms execute the transformation of Eqs. (5) and (6), and then their features are concatenated, resulting in the following output feature representation:

$$\mathbf{Z}_u^{l+1} = concate\left(\sum_{v\in\boldsymbol{N}_u} \mathbf{S}_{u,v}^{l,1}\mathbf{B}_v^{l,1}, \ldots, \sum_{v\in\boldsymbol{N}_u} \mathbf{S}_{u,v}^{l,k}\mathbf{B}_v^{l,k}\right), \tag{7}$$

$$\mathbf{S}_{u,v}^{l,k} = \sigma\left(\boldsymbol{att}_{u,v} * concate(\mathbf{B}_u^{l,k}, \mathbf{B}_v^{l,k})\right), \tag{8}$$

where we cut the $\mathbf{B}_v^l$ into K slices and $\mathbf{B}_v^{l,k} \in \mathbb{R}^{H\times W\times \frac{C}{K}}$ is the kth slice of $\mathbf{B}_v^l$. $\mathbf{S}_{u,v}^{l,k}$ is the attention map computed by kth attention mechanism. The aggregation process based on multi-head mechanism is illustrated in Fig. 3(b). In all our experiments, we set $K = 2$.

By focusing on the local receptive field, PGCN mentioned above focuses on the relationships between the local areas from the same position of different feature maps. Therefore, we named it as Local PGCN (L-PGCN).

3.3. Non-local PGCN

In this section, we introduce another type of PGCN which focuses on capturing long-range relationships between key points by using non-local operation [36] between feature maps of nodes. For node u and its neighborhood v, the node features $\mathbf{Z}_u^l$ and $\mathbf{Z}_v^l$ are

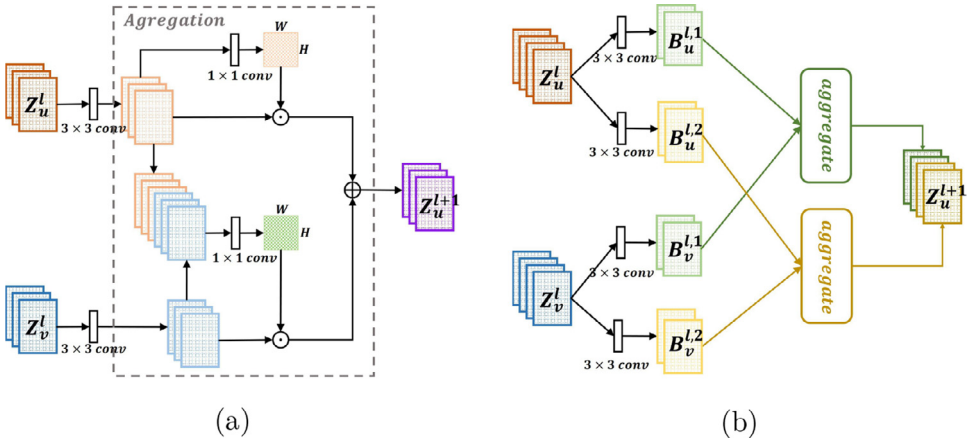

Fig. 3. (a) Layer-wise propagation rule of Local PGCN with single-head attention. Node u is updated by attending over itself and its neighbor v. (b) Illustration of multi-head attention (with $K = 2$ heads) by node u on its adjacent node v. Different colors denote independent attention computations. The aggregated features from each head are concatenated to obtain $\mathbf{Z}_u^{l+1}$. "$\odot$" denotes the channel-wise Hadamard matrix product. "$\oplus$" is element-wise addition.

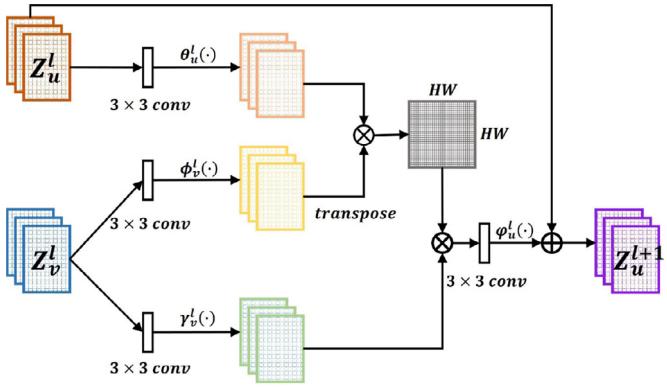

Fig. 4. Layer-wise propagation rule of Non-Local PGCN. "$\otimes$" denotes matrix multiplication. "$\oplus$" is element-wise addition.

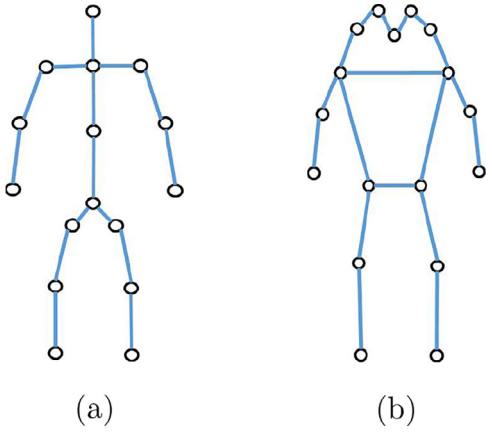

Fig. 5. Natural compositional model of a human body originally presented by (a) MPII dataset and (b) COCO dataset.

first transformed into two feature space $\theta_u^l(\cdot)$, $\phi_v^l(\cdot)$ respectively to calculate the attention $\boldsymbol{\beta}_{u,v} \in \mathbb{R}^{HW \times HW}$:

$$\boldsymbol{\beta}_{u,v} = \frac{1}{HW}\theta_u^l(\mathbf{Z}_u^l)\phi_v^l(\mathbf{Z}_v^l)^\top, \tag{9}$$

where each element of $\boldsymbol{\beta}_{u,v}$ denoted as $\boldsymbol{\beta}_{u,v}(j, i)$ indicates the extent to which the model attends to the i_{th} location of node v when generates the jth feature of node u. Then the output of node u is $\mathbf{Z}_u^{l+1} \in \mathbb{R}^{H \times W \times C}$:

$$\mathbf{Z}_u^{l+1} = \mathbf{Z}_u^l + \varphi_u^l\left(\sum_{v \in \mathbf{N}_u} \boldsymbol{\beta}_{u,v}\gamma_v^l(\mathbf{Z}_v^l)\right), \tag{10}$$

where $\theta_u^l(\cdot)$, $\phi_v^l(\cdot)$, $\gamma_v^l(\cdot)$ and $\varphi_u^l(\cdot)$ are 3×3 convolutions. Here multi-head attention is not used. Due to non-local operation, we name this PGCN as Non-Local PGCN (NL-PGCN). The layer-wise propagation rule is illustrated in Fig. 4.

3.4. Learning

As shown in Fig. 2, The overall model employs a Local PGCN module and a Non-Local PGCN module in parallel to capture local and long-range relationships between key points. 3×3 convolutions are applied on output feature maps of every nodes (key points) to generate the heat maps. We denote the heat maps generated by L-PGCN and NL-PGCN as $\mathbf{P}_u^L$ and $\mathbf{P}_u^{NL}$ respectively. Each of them undergoes a 5×5 convolutions and are then added together. The final heat map $\mathbf{P}_u$ is generated by applying another 3×3 convolutions on the added feature.

$$\mathbf{P}_u = f_{conv3\times3}\left(f_{conv5\times5}(\mathbf{P}_u^L) + f_{conv5\times5}(\mathbf{P}_u^{NL})\right). \tag{11}$$

The ℓ_2 loss is enforced to penalize the difference between predicted heat maps and ground-truth heat maps:

$$l_m = \sum_{u \in \mathcal{G}} ||\mathbf{P}_u^L - \mathbf{G}_u||_2 + ||\mathbf{P}_u^{NL} - \mathbf{G}_u||_2 + ||\mathbf{P}_u - \mathbf{G}_u||_2, \tag{12}$$

where $\mathbf{G}_u$ represents the ground-truth heat map for key point u.

In standard dataset, the ground-truth poses are provided as the key points locations. Denote the ground-truth locations of key point u by (x_u, y_u). Then the ground-truth heat map $\mathbf{G}_u$ of key point u is generated by using a 2D Gaussian centered at (x_u, y_u).

3.5. Backbone

In order to prove the generality of our PGCN model, we place it on top of two different backbone networks. The one is ResNet [37] based network, the other is widely used stacked hourglass network [7].

3.5.1. ResNet

ResNet [37] is the most common backbone network for image feature extraction. Here, we describe our backbone network structure based on the ResNet [37]. We denote the feature activations

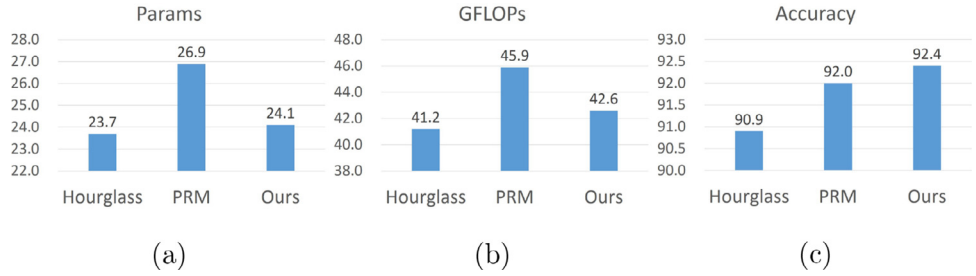

Fig. 6. Statistics of (a) accuracy, (b) number of parameters, and (c) computational complexity in terms of GFLOPs on three models, i.e., Hourglass, PRM, our model.

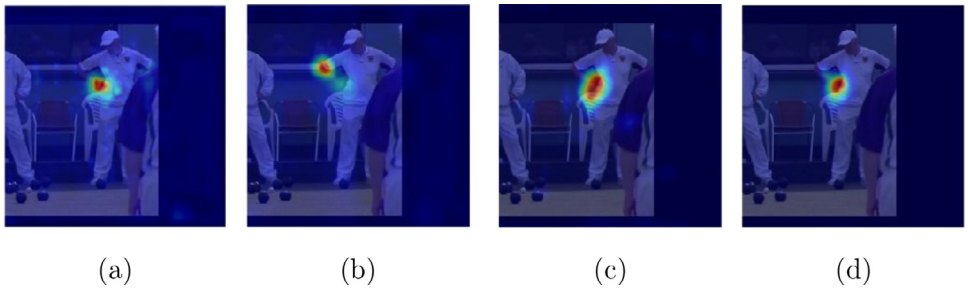

Fig. 7. Feature maps of (a) right wrist, (b) right elbow before spatial attention produced by L-PGCN. Feature maps of right elbow are activated at the location of right wrist, thus information could be propagated from right elbow to right wrist. (c) Attention map from right wrist to right elbow. (d) Feature maps of right elbow after weighted by the attention map (c). Crucial information from right elbow to right wrist is attended.

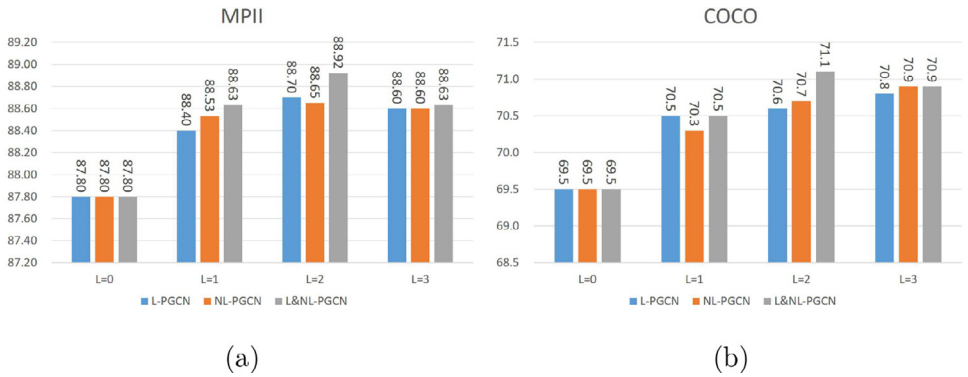

Fig. 8. Comparisons with different depths of PGCN in our mode on (a) MPII validation set and (b) COCO val2017 dataset.

output by each stage's last residual block as {*C*2, *C*3, *C*4, *C*5} respectively. These features are either too coarse to localization or too low-level to recognition. Thus a U-shape structure is integrated to produce a single high-level feature maps of a fine resolution on which the predictions are to be made. Although FPN further improves the U-shape structure with deeply supervised information on detection task, we found it is useless or even harmful for pose estimation. Therefore, we apply a U-shape structure for pose estimation. Specifically, bilinear upsampling followed by a 1×1 convolution is used to upsample spatially coarse but semantically strong feature maps from higher pyramid levels in top-down pathway. Then features from bottom-up and top-down pathways are merged via lateral connections which are 1×1 convolutions. Once obtained, the final feature maps are used to generate initial feature description $\mathbf{Z}^0$ of each key point which is then fed into our L-PGCN and NL-PGCN modules.

3.5.2. Hourglass

The 8-stack Hourglass network is a widely used network framework in single-human pose estimation. In each hourglass stack, features are pooled down to a very low resolution, then they are upsampled and combined with high-resolution features. This structure is repeated for several times to gradually capture more global representations. Equipped with our proposed PGCN, state-of-the-art results is achieved on the pose estimation benchmark datasets.

4. Experiments

In this section, we first describe empirical settings with implementation details. Then, we report the comparison results on both single- and multi-person benchmark datasets. In the following, ablation studies and visualization analyses are presented.

4.1. Empirical settings

Datasets, evaluation metrics and implementation details are presented in this section.

4.1.1. Datasets and evaluation protocols

For single-person pose estimation, we conduct experiments on the MPII [38] and extended LSP [39] datasets. The extended LSP dataset [39] consists of 11k training images and 1k testing images of mostly sports people. The images have been scaled such that the most prominent person is roughly 150 pixels in length. Each image has been annotated with 14 key point locations. Left and right key points are consistently labeled from a person-centric viewpoint.

Fig. 9. (a) Heat map on MPII valid set produced by the ResNet-50 baseline, ResNet-50+L-PGCN and Ground-Truth. (b) Heat map on MPII valid set produced by the ResNet-50 baseline, ResNet-50+NL-PGCN and Ground-Truth.

The MPII Human Pose dataset [38] is a benchmark for evaluation of articulated human pose estimation. The dataset includes around 25k images containing over 40k people with annotated body key points (28k training and 11k testing). Following [7], 3k samples are taken as a validation set to tune the hyper-parameters. For these two datasets, standard Percentage of Correct Key points (PCK) metric is used for evaluation. It reports the percentage of key points that fall into a normalized distance of the ground-truth. For LSP, distance is normalized by torso size, and for MPII, distance is normalized by head size. MPII evaluation metric is referred to PCKh.

For multi-person pose estimation, the COCO Key point Challenge [40] requires localization of multi-person key points in challenging uncontrolled conditions. The key point task involves simultaneously detecting people and localizing their key points (person locations are not given at test time). The COCO train, validation, and test sets contain more than 200k images and 250k person instances labeled with key points. 150k instances of them are publicly available for training and validation. Our models are only trained on COCO train2017 dataset (includes 57k images and 150k person instances) with no extra data involved and validated on the val2017 set (includes 5000 images). The COCO evaluation defines the object key point similarity (OKS) which plays the same role as the IoU. It is calculated from the Euclidean distance between predicted points and ground-truth points normalized by scale of the person and variation of human annotations. Then the mean average precision (AP) over 10 OKS thresholds as main competition metric.

4.1.2. Implementation details

Our network is implemented by using the open-source library PyTorch. For experimental details, we employ Adam [41] with a learning rate 0.001 as the network optimizer. We drop the learning rate by a factor of 10 at the 90th and 110th epochs. Training ends at 140 epochs. The ResNet backbone network is initialized with weight of public-released ImageNet [42] pre-trained model and the rest of our model is randomly initialized.

For the Non-Local PGCN, we do not obtain the attention map $\boldsymbol{\beta}_{u,v} \in \mathbb{R}^{HW \times HW}$ directly, which significantly reduces the computational complexity. Layer-wise propagation rule of Non-Local PGCN can be written as:

$$\mathbf{Z}_u^{l+1} = \mathbf{Z}_u^l + \varphi_u^l\left(\sum_{v \in \boldsymbol{N}_u} \frac{1}{HW}\theta_u^l(\mathbf{Z}_u^l)\phi_v^l(\mathbf{Z}_v^l)^\top \gamma_v^l(\mathbf{Z}_v^l)\right), \tag{13}$$

where $\theta_u^l(\mathbf{Z}_u^l), \phi_v^l(\mathbf{Z}_v^l), \gamma_v^l(\mathbf{Z}_v^l) \in \mathbb{R}^{HW \times C}$ are the transformed feature maps. We implement Eq. (13) by fist calculating $\mathbf{Kernel}_v \in \mathbb{R}^{C \times C}$:

$$\mathbf{Kernel}_v = \frac{1}{HW}\phi_v^l(\mathbf{Z}_v^l)^\top \gamma_v^l(\mathbf{Z}_v^l)\,. \tag{14}$$

Then, Eq. (13) can be rewritten as:

$$\mathbf{Z}_u^{l+1} = \mathbf{Z}_u^l + \varphi_u^l\left(\theta_u^l(\mathbf{Z}_u^l)\sum_{v \in \boldsymbol{N}_u} \mathbf{Kernel}_v\right). \tag{15}$$

Therefore, each layer Non-Local PGCN contains $2J$ matrix multiplication and each matrix multiplication requires HWC^2 FLOPs. J is the number of key points. For the MPII dataset, we have $H = 64, W = 64, C = 16$ and $J = 16$. The matrix multiplication in each layer Non-local PGCN requires $2JHWC^2 = 0.031$ GFLOPs, which is a slight computational burden.

MPII and LSP: For MPII, the scale and position are provided. We first utilize these value to crop the image around the target person and then resize the cropped image to 256 × 256. Data augmentation includes random flip, random rotation (−30, 30) and random scale (0.75, 1.25). Following [43], we estimate the scale and position according to key point positions or image sizes for LSP dataset. For the LSP test set, we perform similar resizing and cropping (or padding), but simply use the image center as the body position, and estimate the body scale by the image size. The compositional model of the human body originally presented by MPII and LSP is shown in Fig. 5(a).

Testing is conducted on six-scale image pyramids with flipping where scale ranges from 0.8 to 1.3 with step of 0.1. For each scale, we run both original input and a flipped version of it and average the heat maps together. Then we warp the heat maps of each scale to original image size and average them to get final heat maps. A quarter of a pixel offset in the direction from the highest response to its next highest neighbor is used to obtain the final location of the key points.

MSCOCO: Following [26], Xiao et al. [27], each ground-truth human box is extended to fixed aspect ratio, e.g., height : width = 4 : 3 and enlarged to contain more context by a rescale factor 1.25. Then the resulting box is cropped from image without distorting image aspect ratio and resized to a fixed resolution. The default resolution is 256 : 192. After cropping from images, we apply random flip, random rotation (-40, 40) and random scale (0.7, 1.3). The natural compositional model of a human body originally presented by COCO is shown in Fig. 5(b).

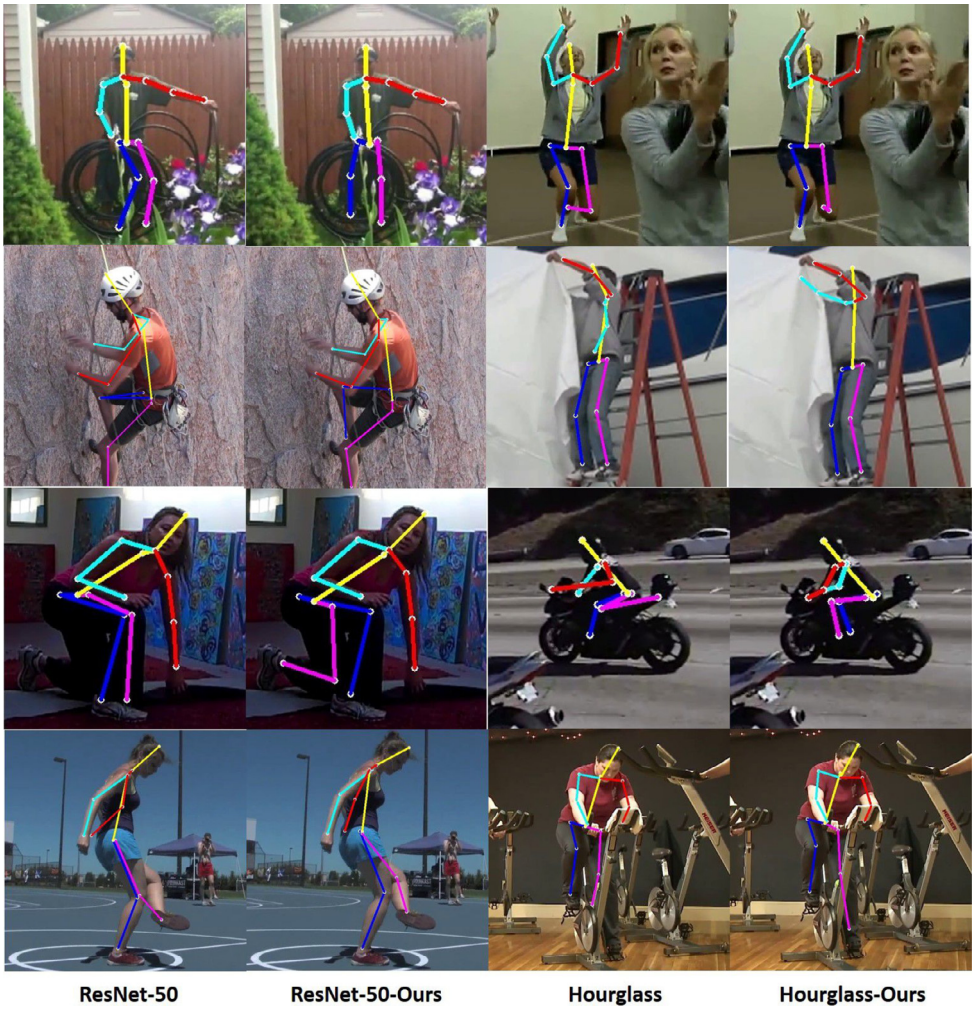

Fig. 10. Prediction samples on the MPII test set produced by different networks, i.e., ResNet-50, ResNet-50-Ours, Hourglass, Hourglass-Ours.

We use the human detection results provided by Chen et al. [26], which achieves detection AP 55.3 for human category in COCO val2017. We also predict the pose of the corresponding flipped image and average the heat maps to get the final prediction. To improve performance at high precision thresholds the prediction is offset by a quarter of a pixel in the direction of its next highest neighbor before transforming back to the original coordinate space of the image.

4.2. Main results

We report the empirical results and comparisons on both single-person and multi-person pose estimation benchmark datasets.

4.2.1. Single-person pose estimation

Results on the MPII dataset: Table 1 summarizes the MPII evaluation results. Ours-Hourglass and Ours-ResNet-50 denote a 8-staked Hourglass backbone and a ResNet-50 based backbone network combined with our L&NL-PGCN model. They are trained on all the MPII training set. We can observe that Ours-Hourglass achieves 92.4% PCKh score at threshold of 0.5, which is the new state-of-the-art result. In particular, it achieves 1.9% and 2.5% improvements on wrist and elbow which are considered as the most challenging key points to be detected. It is noteworthy that Ours-ResNet-50 performs better than many deeper network which demonstrates effectiveness of our L&NL-PGCN model.

For model complexity, as shown in Fig. 6, PRM [43] model increases the number of parameters by 1.6% from 23.7 M to 24.1 M given an 8-stacked Hourglass network. while ours only introduces 0.8% extra parameters. In the other hand, GFLOPS of our model is also 7.1% less than PRM for a 256 $\times$ 256 input RGB image. Our model is both effective and efficient.

Results on the LSP dataset: Table 2 presents the PCK scores at the threshold of 0.2. We follow previous methods [6,43] to train our model by adding the MPII training set to the extended LSP training set with person-centric annotations. Our hourglass based model outperforms state-of-art methods by a large margin.

4.2.2. Multi-person pose estimation

Results on COCO val2017: Table 3 compares our model with Hourglass [7], CPN [26] and SIM [27] on COCO val2017 dataset. All the methods use standard top-down paradigm which sequentially performs human box detection and single-person pose estimation. Our model, Hourglass [7] and CPN [26] use the same human detector with the person detection AP 55.3% which is slightly lower than SIM's 56.4%.

Compared with Hourglass [7,26], our model achieves an improvement of 4.2% in AP. Both methods use an input size 256 $\times$ 192. CPN [26], SIM [27] and our model use the same backbone of ResNet-50. Our model outperforms CPN [26] and SIM [27] by 1.7% and 0.7% for input size 256 $\times$ 192 respectively. When input size is 384 $\times$ 288, our model is better than CPN [26] and SIM [27] by 1.3% and 0.7% AP.

Table 1
Comparisons of PCKh@0.5 scores on the MPII testing set.

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Tompson et al. [8]	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6
Tompson et al. [44]	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
Pishchulin et al.[17]	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Lifshitz et al. [45]	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0
Rafi et al. [46]	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3
Insafutdinov et al. [47]	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
Wei et al. [6]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Bulat and Tzimiropoulos [48]	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Chu et al. [49]	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
Chen et al. [16]	98.1	96.5	92.5	88.5	90.2	**89.6**	86.0	91.9
Yang et al. [43]	***98.5***	96.7	92.5	88.7	91.1	88.6	86.0	92.0
Newell et al. [7]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
ResNet-50-Ours	97.9	96.1	91.5	86.8	90.7	87.6	84.3	91.1
Hourglass-Ours	98.0	***96.9***	***92.7***	***89.0***	***91.8***	*89.4*	***86.1***	***92.4***

Table 2
Comparisons of PCK@0.2 scores on the LSP testing set.

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Lifshitz et al. [45]	96.8	89.0	82.7	79.1	90.9	86.0	82.5	86.7
Pishchulin et al. [17]	97.0	91.0	83.8	78.1	91.0	86.7	82.0	87.1
Insafutdinov et al. [47]	97.4	92.7	87.5	84.4	91.5	89.9	87.2	90.1
Wei et al. [6]	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5
Bulat and Tzimiropoulos [48]	97.2	92.1	88.1	85.2	92.2	91.4	88.7	90.7
Chu et al. [49]	98.1	93.7	89.3	86.9	93.4	94.0	92.5	92.6
Chen et al. [16]	98.5	94.0	89.8	87.5	93.9	94.1	93.0	93.1
Yang et al. [43]	98.3	94.5	92.2	88.9	94.4	95.0	93.7	93.9
Tang et al. [15]	98.3	95.9	93.5	90.7	95.0	**96.6**	95.7	95.1
ResNet-50-Ours	98.6	95.6	94.8	92.9	94.6	96.0	94.7	95.4
Hourglass-Ours	***98.6***	***96.2***	***95.3***	***95.1***	***95.2***	*96.3*	***96.0***	***96.1***

Table 3
Comparison with Hourglass [7], CPN [26] and SIM [27] on COCO val2017 dataset. Their results are cited from [26,27].

Method	Backbone	Input size	$AP_{HumanBox}$	Params	GFLOPs	AP
8stage Hourglass [7]	–	256 × 192	55.3	25.1M	14.3	66.9
CPN [26]	ResNet-50	256 × 192	55.3	27.0M	6.20	69.4
CPN [26]	ResNet-50	384 × 288	55.3	27.0M	13.9	71.6
SIM [27]	ResNet-50	256 × 192	56.4	34.0M	8.9	70.4
SIM [27]	ResNet-50	384 × 288	56.4	34.0M	20.0	72.2
ResNet-50-Ours	ResNet-50	256 × 192	55.3	24.8M	4.6	69.5
Ours	ResNet-50	256 × 192	55.3	25.2M	5.7	***71.1***
Ours	ResNet-50	384 × 288	55.3	25.2M	12.9	***72.9***

4.3. Ablation studies

In this section, we conduct the ablation studies both on single- and multi-person pose estimation task, using the validation set of MPII and COCO datasets respectively. A ResNet-50 based U-shape network is used as the baseline model which achieves a PCKh score at 87.8% on MPII validation set and an AP of 69.5% on COCO val2017 dataset. The overall results are shown in Table 5. Based on the baseline network, we analyze each component of our model, i.e., the local PGCN (L-PGCN) module and the Non-Local PGCN (NL-PGCN) module, by comparing the PCKh score at threshold 0.5 on MPII validation set and AP on COCO val2017 dataset.

4.3.1. Local PGCN

We first evaluate the effect of Local PGCN (L-PGCN) module. By adding our L-PGCN module at the end of the baseline model, we get an PCKh score at 88.6%, which is about 0.8% higher than the baseline model. The AP of the baseline model is improved from 69.5% to 70.6%. The results validate the effectiveness of our L-PGCN module.

4.3.2. Non-local PGCN

We are also interested in how Non-Local PGCN (NL-PGCN) module perform solely. To this end, we conduct an experiment by adding a NL-PGCN module at the end of the baseline model. The PCKh score and AP here is 88.6% and 70.6%, which is 0.8% improvement on PCKh score and 1.2% improvement on AP brought by NL-PGCN.

4.3.3. L&NL-PGCN

While L-PGCN module pays attention to correlation in local area, the NL-PGCN module focus on long-range relations. Then our final PGCN model (L&NL-PGCN) employs a NL-PGCN module and a L-PGCN module in parallel on top of the baseline model. This improves PCKh score from 88.7% to 88.9% and AP from 70.7% to 71.1%, which indicates that our L-PGCN module and NL-PGCN module are complementary with each other.

We also conduct an experiment to explore relationship between the number of PGCN layer and the system performance. As shown in Fig. 8, when the number of graph convolution layers increases, the pose estimation performance increases and saturates quickly at $L = 3$ (both L-PGCN and NL-PGCN). On MPII dataset, PCKh score

Fig. 11. Examples of estimated poses on the MPII test set and the LSP test set.

increases 0.6% and 0.3% when a Local pose graph convolution layer is added incrementally upon our ResNet-50 baseline. Then performance begins to drop at $L = 3$. NL-PGCN has almost the same behavior. The similar trends are found on COCO dataset. The tiny difference is caused by the slightly different graph of human body structure presented by MPII and COCO datasets. It is indicated that relationships between directly adjacent key points are of most importance. The performance drop with increasing number of layers indicates that it is hard to capture and exploit relations between key points in distance.

The motivation of our L-PGCN module is to capture the structural relationships between body key points by propagating information between the local areas of adjacent key points. In order to clearly explain the mechanism of L-PGCN and validate its effectiveness, we visualize the feature maps and attention maps of L-PGCN produced by ResNet-50 + L&NL-PGCN model.

Each layer of L-PGCN first transforms the input key points feature maps into a higher-level feature space by convolutions, which contains both information of the key point itself and its adjacent key points. As shown in the Fig. 7(a,b), feature maps of right elbow are activated at the location of both right wrist and right elbow. As a result, right wrist could take information from feature maps of right elbow.

Then, L-PGCN generates a spatial attention map to focus on crucial information between key points. As shown in the Fig. 7(c), the attention map from right wrist to right elbow is activated around the location of right wrist. Furthermore, weighted by the attention map, feature maps of right elbow are converted to highly localized feature maps of right wrist. These qualitative results demonstrate that L-PGCN learns to propagate structural information between adjacent key point for accurate key point localization.

Table 4
Ablation study of the adjacent matrix on MPII validation set.

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
L-PGCN-I	96.1	95.0	88.4	83.0	87.9	83.9	78.7	88.2
L-PGCN	***96.3***	***95.2***	***89.1***	***83.9***	***88.6***	***84.5***	***79.9***	***88.7***

We also conduct experiments to investigate the importance of structural relationships between key points on MPII validation set. The L-PGCN model is used as the baseline, which achieves a PCKh score at 88.7% on MPII validation set. By remove the feature aggregation process between key points, L-PGCN model is degraded to grouped convolution with spatial attention, which has almost the same capacity as L-PGCN. We call this degraded model as L-PGCN-I, for its adjacent matrix is an identity matrix. As shown in Table 4, L-PGCN-I gets an PCKh score at 88.2%, which is about 0.5% lower than the baseline L-PGCN. The results validate that our L-PGCN improves the localization performance by modeling structural relationships between key points.

4.4. Qualitative results

Fig. 9(a) visualizes some heat maps produced by the ResNet-50 baseline model, ResNet-50 + L-PGCN model and ground-truth. We can observe that our L-PGCN predicts more refined heat maps than the baseline model. Especially when symmetric key points are very close or even overlap, L-PGCN refines the heat map by reducing the ambiguities via utilizing the correlations in the local area.

Table 5
Ablation study on MPII and COCO validation set.

Model	L-PGCN	NL-PGCN	MPII			COCO	
			Wri.	Ank.	Mean	AP	AR
ResNet-50			82.6	78.6	87.8	69.5	75.3
ResNet-50 + L-PGCN	✓		83.6	79.9	88.7	70.6	76.3
ResNet-50 + NL-PGCN		✓	83.3	80.5	88.6	70.7	76.5
ResNet-50 + L&NL-PGCN	✓	✓	***83.6***	***80.8***	***88.9***	***71.1***	***77.0***

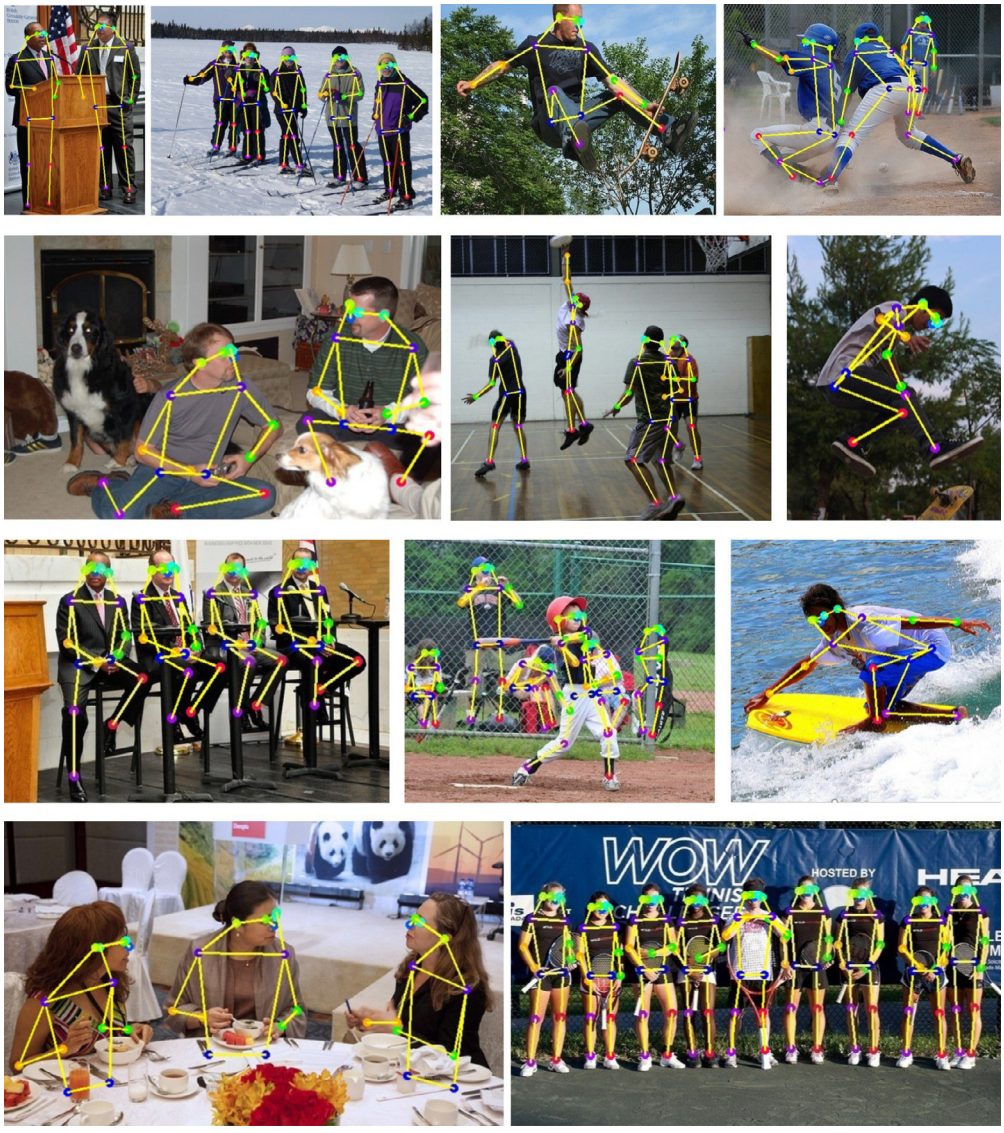

Fig. 12. Examples of estimated poses on the COCO test set.

Fig. 9(b) displays some heat maps produced by the ResNet-50 baseline model, ResNet-50 + NL-PGCN model and ground truth. It is clear to see that our NL-PGCN can associate occluded key points and distinguish the symmetric key points which are widely separated in space.

Fig. 10 displays some pose estimation results obtained by the baseline model and our approach. We observe the baseline model may has difficulty in distinguishing objects with similar appearance with limbs (e.g., the exercise bike in {*col*.3, *row*.4}), and reasoning occluded key points (e.g., the occluded right wrist in {*col*.3, *row*.2}). Our PGCN model would be great help for resolving the ambiguities and occlusions by utilizing the feature response of the neighborhood key points. Fig. 11 demonstrates the poses predicted by our model on the MPII test set and the LSP test set. Our model is robust to extremely difficult cases, e.g., rare poses and cluttered background. More results on COCO test-dev dataset, generated using our model, are shown in Fig. 12.

5. Conclusion

Capturing structured relations between human body key points is the crucial issue for pose estimation. In this paper, we proposed a novel pose estimation model, which leveraged the power of graph convolutional networks to explicitly model the structured

relationships between key points. We built a directed graph between body key points according to body structure where each node (key point) was represented by a tensor which was initially generated by backbone models and attention mechanism was further used to focus on crucial edges (structural information). Then, PGCN mapped this key points graph to a set of structure-aware key points representations. Furthermore, we designed two modules for the model, i.e., Local PGCN and Non-Local PGCN which were proposed to refine the location of key points locally, and capture global underlying contextual information, respectively. Equipped with these two modules, both quantitative results and qualitative visualization validated the effectiveness of our proposed model. Beyond that, extending the model to simultaneously incorporate structural information between multiple persons is an interesting future work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under grant 61871435 and the Fundamental Research Funds for the Central Universities no. 2019kfyXKJC024.

References

[1] L.L. Presti, M. La Cascia, 3D skeleton-based human action classification: a survey, Pattern Recognit. 53 (2016) 130–147.
[2] Y. Guo, Y. Li, Z. Shao, DSRF: a flexible trajectory descriptor for articulated human action recognition, Pattern Recognit. 76 (2018) 137–148.
[3] L. Huang, Y. Huang, W. Ouyang, L. Wang, Part-aligned pose-guided recurrent network for action recognition, Pattern Recognit. 92 (2019) 165–176.
[4] E.P. Ijjina, K.M. C, Classification of human actions using pose-based features and stacked auto encoder, Pattern Recognit. Lett. 83 (2016) 268–277.
[5] H. Wang, L. Wang, Learning content and style: joint action recognition and person identification from human skeletons, Pattern Recognit. 81 (2018) 23–35.
[6] S.-E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, Convolutional pose machines, in: CVPR, 2016, pp. 4724–4732.
[7] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: ECCV, 2016, pp. 483–499.
[8] J.J. Tompson, A. Jain, Y. LeCun, C. Bregler, Joint training of a convolutional network and a graphical model for human pose estimation, in: NIPS, 2014, pp. 1799–1807.
[9] W. Yang, W. Ouyang, H. Li, X. Wang, End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation, in: CVPR, 2016, pp. 3073–3082.
[10] P.F. Felzenszwalb, D.P. Huttenlocher, Pictorial structures for object recognition, IJCV 61 (1) (2005) 55–79.
[11] J. Darby, B. Li, N. Costen, Tracking human pose with multiple activity models, Pattern Recognit. 43 (9) (2010) 3042–3058.
[12] S. Sedai, M. Bennamoun, D.Q. Huynh, Discriminative fusion of shape and appearance features for human pose estimation, Pattern Recognit. 46 (12) (2013) 3223–3237.
[13] A. Toshev, C. Szegedy, DeepPose: human pose estimation via deep neural networks, in: CVPR, 2014, pp. 1653–1660.
[14] X. Chu, W. Ouyang, H. Li, X. Wang, Structured feature learning for pose estimation, in: CVPR, 2016, pp. 4715–4723.
[15] W. Tang, P. Yu, Y. Wu, Deeply learned compositional models for human pose estimation, in: ECCV, 2018, pp. 190–206.
[16] Y. Chen, C. Shen, X.-S. Wei, L. Liu, J. Yang, Adversarial PoseNet: a structure-aware convolutional network for human pose estimation, in: ICCV, 2017, pp. 1212–1221.
[17] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P.V. Gehler, B. Schiele, DeepCut: joint subset partition and labeling for multi person pose estimation, in: CVPR, 2016, pp. 4929–4937.
[18] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2D pose estimation using part affinity fields, in: CVPR, 2017, pp. 7291–7299.
[19] M. Li, Z. Zhou, J. Li, X. Liu, Bottom-up pose estimation of multiple person with bounding box constraint, in: ICPR, 2018, pp. 115–120.
[20] M. Kocabas, S. Karagoz, E. Akbas, MultiPoseNet: fast multi-person pose estimation using pose residual network, in: ECCV, 2018, pp. 417–433.
[21] A. Newell, Z. Huang, J. Deng, Associative embedding: end-to-end learning for joint detection and grouping, in: NIPS, 2017, pp. 2277–2287.
[22] Y. Zhao, Z. Luo, C. Quan, D. Liu, G. Wang, Cluster-wise learning network for multi-person pose estimation, Pattern Recognit. 98 (2020) 107074.
[23] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, K. Murphy, Towards accurate multi-person pose estimation in the wild, in: CVPR, 2017, pp. 4903–4911.
[24] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: ICCV, 2017, pp. 2980–2988.
[25] H.-S. Fang, S. Xie, Y.-W. Tai, C. Lu, RMPE: regional multi-person pose estimation, in: ICCV, 2017, pp. 2334–2343.
[26] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, J. Sun, Cascaded pyramid network for multi-person pose estimation, in: CVPR, 2018, pp. 7103–7112.
[27] B. Xiao, H. Wu, Y. Wei, Simple baselines for human pose estimation and tracking, in: ECCV, 2018, pp. 466–481.
[28] J. Bruna, W. Zaremba, A. Szlam, Y. LeCun, Spectral networks and locally connected networks on graphs, arXiv:1312.62031–14.
[29] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, in: NIPS, 2017, pp. 1024–1034.
[30] Z.-M. Chen, X.-S. Wei, P. Wang, Y. Guo, Multi-label image recognition with graph convolutional networks, in: CVPR, 2019, pp. 5177–5186.
[31] L. Yao, C. Mao, Y. Luo, Graph convolutional networks for text classification, in: AAAI, 33, 2019, pp. 7370–7377.
[32] Q. Zhang, Q. Jin, J. Chang, S. Xiang, C. Pan, Kernel-weighted graph convolutional network: a deep learning approach for traffic forecasting, in: ICPR, 2018, pp. 1018–1023.
[33] T. He, X. Jin, Image emotion distribution learning with graph convolutional networks, in: ACMMM, 2019, pp. 382–390.
[34] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, ICLR, 2017.
[35] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, in: AISTATS, 2011, pp. 315–323.
[36] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: CVPR, 2018, pp. 7794–7803.
[37] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778.
[38] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, 2D human pose estimation: new benchmark and state of the art analysis, in: CVPR, 2014, pp. 3686–3693.
[39] S. Johnson, M. Everingham, Clustered pose and nonlinear appearance models for human pose estimation, in: BMVC, 2, 2010, p. 5.
[40] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, in: ECCV, 2014, pp. 740–755.
[41] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv:1412:6980(2014) 1–15.
[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: CVPR, 2009, pp. 248–255.
[43] W. Yang, S. Li, W. Ouyang, H. Li, X. Wang, Learning feature pyramids for human pose estimation, in: ICCV, 2017, pp. 1281–1290.
[44] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, C. Bregler, Efficient object localization using convolutional networks, in: CVPR, 2015, pp. 648–656.
[45] I. Lifshitz, E. Fetaya, S. Ullman, Human pose estimation using deep consensus voting, in: ECCV, 2016, pp. 246–260.
[46] U. Rafi, B. Leibe, J. Gall, I. Kostrikov, An efficient convolutional network for human pose estimation, in: BMVC, 1, 2016, p. 2.
[47] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, B. Schiele, DeeperCut: a deeper, stronger, and faster multi-person pose estimation model, in: ECCV, 2016, pp. 34–50.
[48] A. Bulat, G. Tzimiropoulos, Human pose estimation via convolutional part heatmap regression, in: ECCV, 2016, pp. 717–732.
[49] X. Chu, W. Yang, W. Ouyang, C. Ma, A.L. Yuille, X. Wang, Multi-context attention for human pose estimation, in: CVPR, 2017, pp. 1831–1840.
[50] X.-S. Wei, J.-H. Luo, J. Wu, Z.-H. Zhou, Selective convolutional descriptor aggregation for fine-grained image retrieval, IEEE TIP 26 (6) (2017) 2868–2881.

Yanrui Bin received the B.S. degree from Huazhong University of Science and Technology in 2017. He is currently pursuing the M.S. degree in Huazhong University of Science and Technology. His research focuses on computer vision and pattern recognition.

Zhao-Min Chen received the B.S. degree from Hunan University and is now a Ph.D. candidate of computer science and technology from Nanjing University. He has published several academic papers on international conferences, such as ICME, CVPR, etc. His research interets are deep learning, computer vision and multi-label image recognition.

Xiu-Shen WEI received his BS degree in computer science, and his Ph.D. degree in computer science and technology from Nanjing University. He is now the Research Director of Megvii Research Nanjing, Megvii Technology, China. He has published more than twenty academic papers on the top-tier international journals and conferences, such as IEEE TPAMI, IEEE TIP, IEEE TNNLS, IEEE TKDE, Machine Learning, CVPR, ICCV, IJCAI, ICDM, ACCV, etc. He achieved the first place in the iNaturalist competition (in association with CVPR 2019), the first place in the Apparent Personality Analysis competition (in association with ECCV 2016) and the first runner-up in the Cultural Event Recognition competition (in association with ICCV 2015) as the team director. He also received the Presidential Special Scholarship (the highest honor for Ph.D. students) in Nanjing University, and received the Outstanding Reviewer Award in CVPR 2017. His research interests are computer vision and machine learning. He has served as a PC member of ICCV, CVPR, ECCV, NIPS, IJCAI, AAAI, etc. He is a member of the IEEE.

Xinya Chen received the B.S. degree from Huazhong University of Science and Technology in 2017. She is currently pursuing the M.S. degree in Huazhong University of Science and Technology. Her research interets include computer vision, pattern recognition, and deep learning.

Changxin Gao received the Ph.D. degree in pattern recognition and intelligent systems from Huazhong University of Science and Technology in 2010. He is currently an associate professor with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology. His research interests are pattern recognition and surveillance video analysis.

Nong Sang graduated from Huazhong University of Science and Technology and received his B.S. degree in computer science and engineering in 1990, M.S. degree in pattern recognition and intelligent control in 1993, and Ph.D. degree in pattern recognition and intelligent systems in 2000. He is currently a professor at the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China. His research interests include object detection and recognition, object tracking, image/video semantic segmentation, intelligent processing and analysis of surveillance videos.