Simple Pose Based Graph Reasoning for Human Pose Estimation

Jia Wang
Huaqiao University
Yanmin Luo (lym@hqu.edu.cn)
Huaqiao University
Guihu Bai
Huaqiao University

Research Article

Keywords: Human pose estimation, dilated convolution, graph convolution, graph reasoning

Posted Date: June 15th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1737952/v1

Simple Pose Based Graph Reasoning for Human Pose Estimation

Jia Wang[1,2], Yanmin Luo[1,2*] and Guihu Bai[1,2]

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

[2]Xiamen Key Laboratory of Computer Vision and Pattern Recognition, Huaqiao University, Xiamen, 361021, PR China.

*Corresponding author(s). E-mail(s): lym@hqu.edu.cn;
Contributing authors: 20014083040@stu.hqu.edu.cn; bgh@stu.hqu.edu.cn;

Abstract

Aiming at the challenge of modeling the relations globally among different body joints. In this paper, we proposed a simple network based on graph reasoning for human pose estimation, which named SP-GRe. We introduce dilated convolution to construct a Dilated Bottleneck Module (DRM), which can enlarge the receptive field and exploit its feature extraction capability. Meanwhile, it can enhance the model's local representations of each key point. In view of the potential advantages of graph-based propagation, we design a Global Graph Reasoning module (GGR) based on graph convolution. The module stores the explicit joints in the graph structure for global relationship reasoning. By aggregating the features of local joints and global graph nodes, GGR enables the accurate location of key points in the interaction between projection and back-projection. Comprehensive experiments demonstrate that the proposed method achieves superior top-down pose estimation results on two benchmark datasets, MSCOCO and MPII. Moreover, SP-GRe demonstrates superior results on human pose estimation over popular human pose estimation networks.

Keywords: Human pose estimation, dilated convolution, graph convolution, graph reasoning.

1. Introduction

Human pose estimation(HPE) is an important problem in the field of computer vision, the purpose is to locate key points (human joints or parts). It is the basis of other related tasks and various advanced visual applications, including video pose estimation[1, 2], pose tracking[3] and human action recognition[4, 5]. It is also the basic tool of video surveillance, human-computer interaction, virtual reality[6, 7, 8].

Deep learning[1, 2, 6, 8, 9] has achieved huge success in multi-domain computer tasks. HPE architecture based on Convolutional Neural Networks (CNNs) [3, 10, 11, 12, 13] have also achieved great advantages in computer vision tasks. However, the receptive field of convolution operation is limited, it can only capture local information. CNNs stack the convolution layers into a depth model to aggregate the rich information of the global background. It is very challenging to integrate the prior information of human structure into the deep CNNs. Therefore, recent approaches consider the structure of joints in the human body. For example [14], in the first stage, a heat map is generated to produce an initial pose, and in the second stage, the initial pose obtained from the heat map is refined by an Image-guided Progressive GCN (IPE-GCN) module. And in [15], heatmap regression network is applied to obtain a rough localization result and a set of proposal guided points. Then, for each guided point, different visual feature is extracted by the localization subnet. These methods are two-stage networks in which human joints are refined in the second stage. Therefore, we hope to infer the relationship between different joints in the early stage of CNN networks.

In this paper, we propose a Dilated Bottleneck Module (DBM) to expand the receptive field and avoid the loss of local location information. The standard convolution operation is reformulated by dilating[16], learning the optimal receptive field and increasing the spatial sampling position. And we use dilated convolution instead of standard convolution. Node-based Graph Convolution Network(GCN)[4, 17, 18] exchanges information between adjacent nodes and extracts the neighborhood features of graph. We design a Global Graph Reasoning module, the relationship between joints is processed by graph convolution, and information interaction between local features and global features is carried out in the process of projection and back-projection.

Our contributions can be summarized as follows:

(1) We proposed a simple network based on graph reasoning for human pose estimation: SP-GRe, which is focusing on the interaction between the features of local joints and global graph nodes.

(2) We introduce dilated convolution to build a dilated bottleneck module and enlarge the receptive field. This module explores the search problem within the dilated domain, learning the optimal receptive field in a variety of dilation patterns, covering all body joints in the image.

(3) We design a global graph reasoning module, which infers the global relationship of human joints by graph convolution, and constructs the local and global information interaction of human key points in the process of projection and back-projection.

2. Related Work

2.1. Multi-person Pose Estimation

Top-down method. The top-down method first detects the human body position on the image, and then performs single pose estimation for each detected human body. Fang et al.[19] designed a new Regional Multi-person Pose Estimation (RMPE) method to improve the performance of HPE in complex scenes. Mask-RCNN proposed by HE.et al.[20] predict human bounding box and human key points at the same time. Chen et al. [10] proposed CPN (Cascaded Pyramid Network), which is divided into two stages: GlobalNet and RefineNet. GlobalNet is responsible for the detection of all key points in the network. RefineNet refines the prediction results of GlobalNet. Li et al. [21] designed a global maximum joint association algorithm to solve the association problem in crowded scenes. SRFNet[22] uses skip connection fuse multi-scale feature which allows the network to improve spatial context. FastNet[23] aims at the problem of developing efficient models for HPE algorithms under computation-limited resources. Qiu et al. [14] developed the occlusion pose estimation and correction (OPEC-Net) module and designed an occlusion pose dataset(OCPose) for crowd pose estimation.

Bottom-up method. The bottom-up method[2, 24, 25, 26] detects all candidate joints by applying a joint detector globally, and then clusters them according to a certain method. Pishchulin et al. [24] proposed the first bottom-up pose estimation method. They perform the estimation tasks by solving the minimum cost multi-cut problem, which model candidate joints as vertices and candidate trunk relationships as edges. Insafutdinov et al.[25] introduced deepercut and improved deepcut to improve performance and speed. Cao et al. [2] proposed part affinity fields (PAFs) and effectively associated key points with individuals in the image using Hungarian bipartite matching algorithm. Luo et al.[27] proposed the joints kinship pattern matching algorithm, which calculates the degree of relatedness (call it Kinship) for adjacent joints.

2.2. Graph Convolution Networks

Recently, the graph methods based on CNN [28, 29, 30] can effectively capture the dependencies among joints. Kipf et al.[28] proposed a graph convolution network (GCN) for semi-supervised classification for the first time. Since then, GCN has been widely used in various tasks. Chen et al. [17] proposed a graph-based global reasoning network and designed a global reasoning unit to infer between disjoint and distant regions. In addition, 2D to 3D pose regression is also a graph prediction problem. Zhao et al.[31] proposed a new 2D to 3D human pose estimation method, which uses SemGCN to realize 2D to 3D human pose regression.

3. Our Methods

We aim to capture local processing and global interaction of human pose estimation. The overall framework of our method is shown in Figure1.

3.1. Dilated Bottleneck module

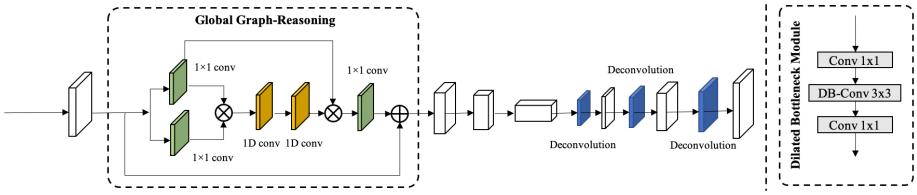
CNNs[32, 33] enlarge the receptive field by stacking multiple convolution layers and down-sampling, restore the original image by up sampling. If the down-sampling ratio under the height and width of feature map is too large, it will be difficult to restore the original image.

To enlarge the receptive field and fully exploit its feature extraction capability, we propose a dilated bottleneck module to enhance the local perception ability of the model for each key point. We use dilated convolution [34] to obtain larger receptive fields and reduce information loss instead of pooling. Based on the optimization method of dilated convolution in [16], this paper explores the search problem in the dilatation domain, carries out independent dilatation between different axes, channels and layers, selects the optimal dilation patterns with the best representation ability.

As shown in Fig2 (b), the pre-trained convolution layer is used to select the optimal expansion mode from a variety of dilation patterns through hybrid dilated convolution, and each layer of the same dilation patterns is recombined to obtain the optimal receptive fields, which is expressed as follows:

$$\mathbb{R} = \left\{ \left(r_x^i, r_y^i \right) \mid r_x^i \in \{1, 2, ..., r_{max}\}, r_y^i \in \{1, 2, ..., r_{max}\} \right\} \tag{1}$$

where $i \in \{1, 2, ..., C\}$, C is the output channel, r_x^i and r_y^i are the dilation values in x axis and y axis of the



(a) Architecture (b) Dilated Bottleneck Module

Fig. 1: The SP-GRe framework. The input image is fed in the backbone, the dilated bottleneck module and the global graph reasoning module generate one heatmap per joint or class. Similar to SimpleBaseline[3], our SP-GRe consists of a backbone network with several deconvolution layers. However, we use the dilated bottleneck module as the bottleneck residual block to enlarge the receptive field and use the global graph-reasoning module to infer the global relationship of joints.

filter at the $i-th$ output channel, ranging from 1 to r_{max}. Calculate the minimum error L1 loss between the output expectation of the pre-training weight W and the output expectation of the sampling expansion weight W_r:

$$min \sum_{i=1}^{C} \omega \left((W^i - W^i_{r^i_x, r^i_y} \cdot Conv(1) \right) \tag{2}$$

where 1 represents an all-ones matrix with the same dimension. The optimal receptive field $\mathbb{R}$ of each convolution layer can be easily solved after traversing all dilation patterns (r^i_x, r^i_y).

We design a dilated bottleneck module instead of bottleneck residual block of ResNet. The standard bottleneck residual block is composed of 1×1 ,3×3 and 1×1convolution. As shown in Figure2(a), the 3×3 convolution in the bottleneck residual block can be easily replaced by an optimized dilated convolution to obtain a new dilated bottleneck module, as shown in Figure2(c).

3.2. Global Graph Reasoning Module

Since convolution operations can only model local relationships, most methods inefficiently build deep networks to capture global relationships between different key points. Inspired by Chen[17] et al., we directly infer the global relation of the structure of human body diagram. We project the local joints features to a global graph space, in which we use graph convolution to achieve global graph reasoning to get the graph node features, and then projected into the downstream coordinate space. Therefore, global graph reasoning can be performed in early stages of human pose estimation. As shown in Figure3.

3.2.1. Local features projection

Firstly, we need to find the projection function that projects the local joint feature to the global graph space, so that the local feature $f \in \mathbb{R}^{W \times H \times C}$ is updated to feature $F \in \mathbb{R}^{N \times C}$ through the projection function $\Lambda(\cdot)$, N is the number of node features in the global graph space. The projection function is expressed as the linear combination of the local features, and the linear projection of the local features is used to obtain new features:

$$F = \Lambda(f) = \lambda_i f = \sum_{\forall j} \lambda_{ij} f_j \tag{3}$$

where $f_j \in \mathbb{R}^{1 \times C}$, the feature f_j of node j is assigned to F weighted by a scalar λ_{ij}. To reduce the input dimension and enhance the capacity of the projection function, learnable projection weights $\lambda = [\lambda_1, \lambda_2, ..., \lambda_n] \in \mathbb{R}^{W \times H \times C}$ can be directly generated by a 1×1 convolution. The resulting projection weight λ is multiplied by the input feature to project the local feature into the global graph space.

3.2.2. Global graph reasoning

Local features projected into the global graph space are regarded as nodes of a fully connected graph. The fully connected graph is inferred by learning edge weights that correspond to interactions of the underlying globally-pooled features of each node. Inspired by the fast graph convolution[16]: $H^{(l+1)} = \sigma\left(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right)$, the graph reasoning in the model is re-expressed as:

$$G = \sigma(\mathbb{L}F\Theta) = \sigma\left(\left(I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}}\right)F\Theta\right) \tag{4}$$

F represents the input features, Θ represents a trainable weight matrix, σ represents the ReLU activation function, and G represents the global node-feature matrix.

Firstly, the graph convolution performs Laplace smoothing to propagate the node features on the graph. During training, the adjacency matrix A learns the weights of edges, which reflect the relationship between the globally-pooled features of each node. After receiving all the necessary information, each node updates the state of the node through the corresponding linear transformation Θ. We use graph convolution via two 1D convolution layers along different directions, channel-wise and node-wise.

3.2.3. Global feature back-projection

To make the above building block compatible with CNN architectures, we project the output features back to the downstream space after the graph reasoning. In this way, the features updated from reasoning can be used by the following convolution layer to make better decisions and obtain the information between key points. The back-projection is similar to projection. Given the node-feature matrix $G \in \mathbb{R}^{N \times C}$, we aim to learn a back-projection function $\Phi(\cdot)$,

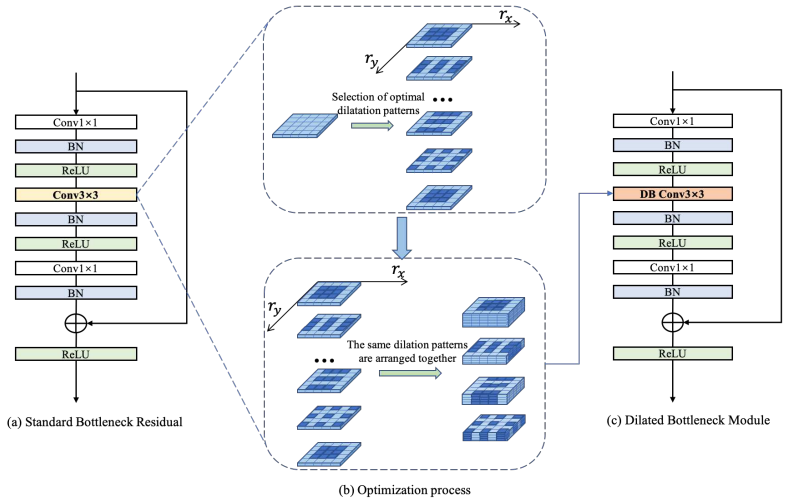


Fig. 2: Dilated Bottleneck Module.

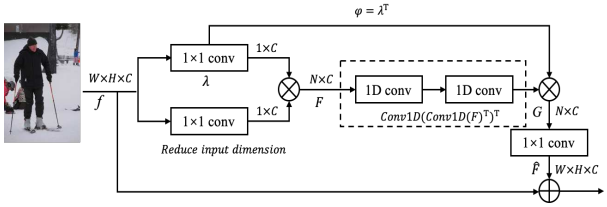


Fig. 3: Global graph reasoning. The local feature $f \in \mathbb{R}^{W\times H\times C}$ is updated to $F \in \mathbb{R}^{N\times C}$ through 1×1 convolution, and then the learnable projection weight λ is generated, while the input dimension is also reduced. The information generated by graph convolution is back-projected to the downstream coordinate space, and the output of the convolution layer is reused as the weight. Another convolution layer is attached after migrating the information back to the downstream coordinate space for dimension expansion, so that the output dimension can match the input dimension forming a residual path.

transform the feature to $\widehat{F} \in \mathbb{R}^{W\times H\times C}$,

$$\widehat{F} = \Phi(G) = \varphi_i G = \sum_{\forall j} \varphi_{ij} g_j \tag{5}$$

similar to the projection, we can set $\varphi = \lambda^\top$ and reuse the projection λ to reduce the computational cost without any negative impact on the final accuracy.

Figure3 shows the detailed process of global graph reasoning. The information of graph convolution is projected back to the original coordinate space through formula5, and the output of convolution is reused as the weight. After the information is migrated back to the original coordinate space for dimension expansion, another convolution layer is added to match the output dimension with the input dimension to form a residual path. Finally, the dilated bottleneck module and the global graph reasoning module are integrated into the ResNet network, followed by three deconvolution and convolution alternating networks to obtain a high-resolution heatmap.

4. Experiments

In order to evaluate our proposed global graph reasoning network, we conducted comprehensive experiments on MPII[35] data set and COCO[36] data set. Our implementation details and data sets are described below.

4.1. Implementation details

For fair comparison, we used the same training configuration as [3], trained 180 epochs on MPII training set and 140 epochs on COCO training set. We use the Adam optimizer, which basic learning rate is le-3. Similar to[37], we use Mean Squared Error (MSE) loss, the target's heatmap is achieved by using a 2D Gaussian distribution near the target location. We used three different input sizes (256×192, 256×256,384×288) and experiment on ResNet50 and ResNet101. We also further experiment on a high-resolution network HRNet.The networks are all trained on 1 Nvidia GTX 3080Ti GPUs.

4.2. COCO Dataset

The COCO dataset provides challenging images with different body poses, different body scales and occlusion patterns in the wild. It contains 200K images and 250K person instances labeled with 17 joints. We trained our model on the COCO train2017 dataset, including 57K images and 150K human instances. We evaluate our method on the val2017 set and test-dev2017 set, which contains 5000 images and 20K images.

4.2.1. Evaluation metric

The standard evaluation index is based on Object Keypoint Similarity metric (OKS): $OKS = \frac{\sum_i exp\left(-\frac{d_i^2}{2s^2k_i^2}\delta\right)(v_i>0)}{\sum_i \delta(v_i>0)}$, d_i represents the Euclidean distance between the detected key point and the corresponding ground truth, v_i represents the visible sign

of the ground truth, s is the scale of the object, k_i controls falloff. We report standard average precision and recall scores: AP^{50}, AP^{75}, AP, AP^M, AP^L and AR.

4.2.2. Results on the validation set

According to the width of the network, our network includes SP-GRe-50 and SP-GRe-101. Table1 reports the pose estimation performances of our method and baseline in ResNet50 and ResNet101, and images input size are 256×192 and 384x288. SP-GRe-50 achieves an AP of 75.5 with the input of 384×288. It outperforms all the backbone approaches. The improvement is 3 points for SP-GRe-50 and 1.5 points for SP-GRe-101 in the input of 256 × 192. In the larger input 384 × 288, the improvement is 1.8 points for SP-GRe-50 and 1.6 points for SP-GRe-101. We can find that the improvement of SP-GRe on the small network is more obvious than on the large network, as shown in Figure4a. We report the results of our method and previous state-of-the-art methods in Table2. Compared with CPN and CPN (ensembles), SP-GRe-50 achieves 3.4 and 2.5 points gain. Compared with GGR- GCN[38], SP-GRe-50 improves AP by 2.9 points.

Meanwhile, we report the results with the global graph reasoning module on the backbone HRNet and compared results with baseline, as shown in Table3. With the input size 256× 192, SP-GRe-W32 achieves 76.5 AP, outperforming HRNet-W32 with the same input size. With the input size 384× 288, SP-GRe-W32 achieve 3.5 points gain. Figure4b shows the accuracy in different image input sizes on COCO for SP-GRe-W32. Figure6 shows the visual results on COCO from SP-GRe-W32.

4.2.3. Results on the test set

Table4 reports the pose estimation performances of our approach and the existing state-of-the-art approaches. Our approach is significantly better than many top-down approaches. On the other hand, our small network, SP-GRe-50, with the input size 256 × 192, achieves an AP of 70.7. It outperforms the baseline. Our big model, SP-GRe-W32, achieves 74.8 AP. The results are also close to the best-reported pose estimation results. Our model using smaller networks ResNet50 has higher accuracy than other models with the same network.

4.3. MPII Dataset

The MPII dataset contains 25K images with over 40K annotated poses, which consists of whole-body annotated images from a wide-range of real-world. There are 12K subjects for testing and the remaining subjects for the training set.

4.3.1. Evaluation metric

The standard metric, the PCKh (head-normalized probability of correct keypoint) score, is used. When the detection of joint is within a threshold distance of the ground truth, the evaluation metric is used to consider whether the prediction of key points is correct. MPII dataset uses threshold of PCKh@0.5, which is 50% of the head diameter.

4.3.2. Results on MPII Dataset

Table5 shows the PCKh@0.5 results on the MPII test set. SP-GRe-W32 achieves a 91.3 PCKh@0.5 score and outperforms the stacked hourglass approach [37]. Our method is not better than the other state-of-the-art methods. There are two reasons: 1) The first might be that the performance in this dataset tends to be saturated. 2) Our model is more suitable for small networks, compared with the big networks, the network expression ability is not enough.

4.4. Ablation Studies

We study the effect of each component in our approach on the COCO val dataset. All results are obtained over the input size of 256 × 192 and 384 × 288. And we use ResNet-50 as the backbone. Figure5 shows how the input image size affects the performance in comparison with SimpleBaseline.

4.4.1. Dilated Bottleneck module

As shown in Table6, the dilated bottleneck module can greatly improve the performance of network. We study the influence of the different image size within dilated bottleneck module in Figure5. For example, the dilated bottleneck module gets 2 AP over the baseline with the input of 256x192, and improves 1.1 points with the input of 384 × 288. It proves the effectiveness of the dilated bottleneck module.

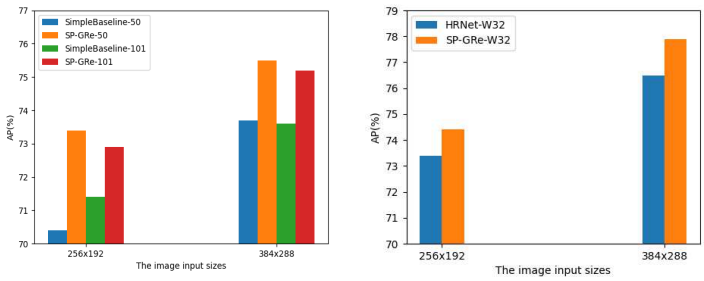
4.4.2. Global Graph Reasoning

As the results from Table6 show, adding the global graph reasoning module lead to a better result with the input of image size 256 × 192. SB-50+GGRM achieves 2.2 points gain. However, we find that the performance decrease with input large image size. This illustrates that global graph reasoning has more advantages in small-scale images. When we add dilated bottleneck module and global graph reasoning module to baseline, the performance is significantly improved in both image sizes 256 × 192 and 384 × 288.

5. Conclusion

In this paper, we proposed a simple network based on graph reasoning for human pose estimation. We propose a dilated bottleneck module to obtain a large receptive field without pooling and reduce the loss of information. And global graph reasoning module performs global reasoning on human joints through the interaction of projection and back-projection. The local joint feature and global graph node feature are extracted to realize the accurate positioning of key points. Because we focus on a simper human pose estimation network, the performance of this method fails to reach the state-of-the-art performance.

Acknowledgments. This work was supported by Natural Science Foundation of Fujian Province, China under grant 2020J01082.



(a) The backbone: Simplebaseline (b) The backbone: HRNet-W32

Fig. 4: Accuracy of different backbones in different image input sizes

Table 1: Comparisons with SimpleBaseline on the COCO validation set.

Method	Backbone	Input	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Simplebaseline-50	ResNet50	256 × 192	70.4	88.6	78.3	67.1	77.2	76.3
SP-GRe-50	ResNet50	256 × 192	**73.4**	**92.6**	**81.5**	**70.6**	**78**	**76.5**
Simplebaseline-50	ResNet50	384 × 288	73.7	91.9	81.1	70.3	80.0	79.0
SP-GRe-50	ResNet50	384 × 288	**75.5**	**92.5**	**82.6**	**72.6**	**80.3**	**78.4**
Simplebaseline-101	ResNet101	256 × 192	71.4	89.3	79.3	68.1	78.1	77.1
SP-GRe-101	ResNet101	256 × 192	**72.9**	**92.4**	**80.3**	**70.1**	76.9	76
Simplebaseline-101	ResNet101	384 × 288	73.6	89.6	80.3	69.9	81.1	79.1
SP-GRe-101	ResNet101	384 × 288	**75.2**	**92.5**	**82.4**	**72.3**	80.0	78.1

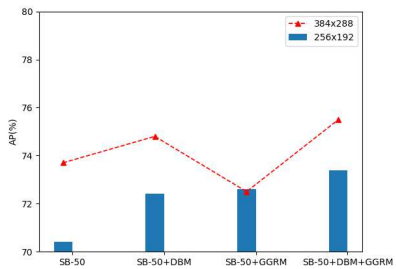


Fig. 5: The input image size affects the performance in comparison with SimpleBaseline.

Fig. 6: Example qualitative results on COCO pose estimation.

Declarations

- Funding Funding The Natural Science Foundation of Fujian Province, China under grant 2020J01082
- Conflicts of interest There are no conflicts of interest.
- Availability of data and materials The data comes from the common dataset
- Code availability Custom code

References

[1] Bruce Xiaohan Nie, Caiming Xiong, and Song-Chun Zhu. Joint action recognition and pose estimation from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1293–1301, 2015.

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[3] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018.

[4] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[5] Chenyang Si, Ya Jing, Wei Wang, Liang Wang, and Tieniu Tan. Skeleton-based action recognition with spatial reasoning and temporal stack learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–118, 2018.

[6] Utkarsh Gaur, Yingying Zhu, Bi Song, and A Roy-Chowdhury. A "string of feature graphs" model for recognition of complex activities in natural videos. In *2011 International conference on computer vision*, pages 2595–2602. IEEE, 2011.

[7] Zoran Duric, Wayne D Gray, Ric Heishman, Fayin Li, Azriel Rosenfeld, Michael J Schoelles, Christian Schunn, and Harry Wechsler. Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction. *Proceedings of the IEEE*, 90(7):1272–1289, 2002.

[8] MR Sudha, K Sriraghav, Shomona Gracia Jacob, S Manisha, et al. Approaches and applications of virtual reality and gesture recognition: A review. *International Journal of Ambient Computing and Intelligence (IJACI)*, 8(4):1–18, 2017.

[9] Xin Liu, Zhikai Hu, Haibin Ling, and Yiu-Ming Cheung. Mtfh: A matrix tri-factorization hashing framework for efficient cross-modal retrieval. *IEEE*

Table 2: Comparisons on the COCO validation set.

Method	Backbone	Input	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
8-stage Hourglass[37]	8-stage Hourglass	256 × 192	66.9	-	-	-	-	-
CPN[10]	ResNet50	256 × 192	68.6	-	-	-	-	-
GRR-GCN[38]	ResNet50	384 × 288	72.6	89.4	79.5	68.7	79.9	78
Simplebaseline-50[3]	ResNet50	256 × 192	70.4	88.6	78.3	67.1	77.2	76.3
Simplebaseline-50[3]	ResNet50	384 × 288	73.7	91.9	81.1	70.3	80.0	79.0
Simplebaseline-101[3]	ResNet101	256 × 192	71.4	89.3	79.3	68.1	78.1	77.1
Simplebaseline-101[3]	ResNet101	384 × 288	73.6	89.6	80.3	69.9	81.1	79.1
SP-GRe-50	ResNet50	256 × 192	**73.4**	**92.6**	**81.5**	**70.6**	**78**	**76.5**
SP-GRe-50	ResNet50	384 × 288	**75.5**	**92.5**	**82.6**	**72.6**	**80.3**	**78.4**
SP-GRe-101	ResNet101	256 × 192	**72.9**	**92.4**	**80.3**	**70.1**	76.9	76
SP-GRe-101	ResNet101	384 × 288	**75.2**	**92.5**	**82.4**	**72.3**	80.0	78.1

Table 3: Comparisons on the COCO validation set.

Method	Backbone	Input	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
HRNet-W32[13]	HRNet-W32	256 × 192	73.4	89.5	80.7	67.1	70.2	78.9
HRNet-W32[13]	HRNet-W32	384 × 288	74.4	90.5	81.9	70.8	81.0	79.8
SP-GRe-W32	HRNet-W32	256 × 192	**76.5**	**93.6**	**73.8**	83.7	81	79.4
SP-GRe-W32	HRNet-W32	384 × 288	**77.9**	**93.6**	**84.7**	**74.8**	**82.4**	**80.4**

Transactions on Pattern Analysis and Machine Intelligence, 43(3):964–981, 2021.

[10] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018.

[11] Wei Yang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3073–3082, 2016.

[12] Kai Su, Dongdong Yu, Zhenqi Xu, Xin Geng, and Changhu Wang. Multi-person pose estimation with enhanced channel-wise and spatial information. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5674–5682, 2019.

[13] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.

[14] Lingteng Qiu, Xuanye Zhang, Yanran Li, Guanbin Li, Xiaojun Wu, Zixiang Xiong, Xiaoguang Han, and Shuguang Cui. Peeking into occluded joints: A novel framework for crowd pose estimation. In *European Conference on Computer Vision*, pages 488–504. Springer, 2020.

[15] Jian Wang, Xiang Long, Yuan Gao, Errui Ding, and Shilei Wen. Graph-pcnn: Two stage human pose estimation with graph pose refinement. In *European Conference on Computer Vision*, pages 492–508. Springer, 2020.

[16] Jie Liu, Chuming Li, Feng Liang, Chen Lin, Ming Sun, Junjie Yan, Wanli Ouyang, and Dong Xu. Inception convolution with efficient dilation search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11486–11495, 2021.

[17] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 433–442, 2019.

[18] Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI conference on artificial intelligence*, 2018.

[19] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2334–2343, 2017.

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[21] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10863–10872, 2019.

[22] Zhilong Ou, YanMin Luo, Jin Chen, and Geng Chen. Srfnet: selective receptive field network for human pose estimation. *The Journal of Supercomputing*, 78(1):691–711, 2022.

[23] Yanmin Luo, Zhilong Ou, Tianjun Wan, and Jing-Ming Guo. Fastnet: Fast high-resolution network for human pose estimation. *Image and Vision Computing*, page 104390, 2022.

[24] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4929–4937, 2016.

[25] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European conference on computer vision*, pages 34–50. Springer, 2016.

[26] Zhiqian Zhang, Yanmin Luo, and Jin Gou. Double anchor embedding for accurate multi-person 2d pose estimation. *Image and Vision Computing*, 111:104198, 2021.

[27] Yanmin Luo, Zhitong Xu, Peizhong Liu, Yongzhao Du, and Jing-Ming Guo. Multi-person pose estimation via multi-layer fractal network and joints kinship pattern. *IEEE Transactions on Image Processing*, 28(1):142–155, 2018.

[28] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[29] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013.

[30] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016.

[31] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings*

Table 4: Comparisons on the COCO test-dev set.

Method	Backbone	Input	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Mask-RCNN[20]	ResNet-50-FPN	-	63.1	87.3	68.7	57.8	71.4	-
G-RMI[39]	ResNet101	353 × 257	64.9	85.5	71.3	62.3	70.0	69.7
G-RMI+extra data[39]	ResNet101	353 × 257	68.5	87.1	75.5	65.8	73.3	73.3
GRR-GCN[38]	ResNet50	384 × 288	71.8	91.1	79.0	68.2	78.3	77.2
RMPE[19]	PyraNet[40]	320 × 256	72.3	89.2	79.1	68.0	78.6	-
SimpleBaseline[3]	ResNet50	256 × 192	70.0	90.9	77.9	66.8	75.8	75.6
CPN[10]	ResNet-Inception	384 × 288	72.1	91.4	80.0	68.7	77.2	78.5
CPN(ensemble)[10]	ResNet-Inception	384 × 288	73.0	91.7	80.9	69.5	78.1	79.0
OPEC-Net[14]	Simple Pose[3]	384 × 288	73.9	91.9	82.2	-	-	-
HRNet-W32[13]	HRNet-W32	384 × 288	74.9	92.5	82.8	71.3	80.9	80.1
SP-GRe-50	ResNet50	256 × 192	70.7	91.0	79.0	67.8	76.2	76.4
SP-GRe-50	ResNet50	384 × 288	71.9	90.9	79.8	68.7	77.7	77.3
SP-GRe-101	ResNet101	384 × 288	73.4	91.6	81.3	70.2	79.4	78.9
SP-GRe-W32	HRNet-W32	256 × 192	73.5	92.2	81.8	70.5	79.2	79.0
SP-GRe-W32	HRNet-W32	384 × 288	74.8	92.4	82.9	71.4	80.7	80.1

Table 5: Performance comparisons on the MPII test set (PCKh@0.5).

Backbone	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Total
Wei et. al.[41]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Bulat et al.[42]	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Newell et al.[37]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Tang et al.[43]	97.4	96.4	92.1	87.7	90.2	87.7	84.3	91.2
Ning et al.[44]	98.1	96.3	92.2	87.8	90.6	87.6	82.7	91.2
Luvizon et al.[45]	98.1	96.6	92.0	87.5	90.6	88.0	82.7	91.2
Chu et al.[46]	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
Chou et al.[47]	98.2	96.8	92.2	88.0	91.3	89.1	84.9	91.8
Yang et al.[40]	98.5	96.7	92.5	88.7	91.1	88.6	86.0	91.9
Ke et al.[48]	98.5	96.8	92.7	88.4	90.6	89.3	86.3	92.1
SimpleBaseline[3]	98.5	96.6	91.9	87.6	91.1	88.1	84.1	91.5
HRNet-W32[13]	98.6	96.9	92.8	89.0	91.5	89.0	85.7	92.3
SP-GRe-50	96.9	94.1	87.6	81.7	86.2	81.6	77.5	87.0
SP-GRe-101	98.3	96.2	91.1	87.2	89.9	87.5	83.5	90.9
SP-GRe-W32	98.3	96.2	91.7	87.6	90.0	87.9	83.9	91.2

Table 6: Ablation about the dilated bottleneck module and global graph reasoning module. SB=SimpleBaseline, DBW= Dilated Bottleneck Module, GGRM= Global Graph-Reasoning Module.

Backbone	Input	AP	AR
SB-50	256 × 192	70.	76.3
SB-50+DBM	256 × 192	72.4	75.5
SB-50+GGRM	256 × 192	72.6	75.7
SB-50+DBM+GGRM	256 × 192	73.4	76.5
SB-50	384 × 288	73.7	79.0
SB-50+DBM	384 × 288	74.8	77.7
SB-50+GGRM	384 × 288	72.5	75.5
SB-50+DBM+GGRM	384 × 288	75.5	78.4

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3425–3435, 2019.

[32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[34] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[35] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[37] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

[38] Rui Wang, Chenyang Huang, and Xiangyang Wang. Global relation reasoning graph convolutional networks for human pose estimation. *IEEE Access*, 8:38472–38480, 2020.

[39] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4903–4911, 2017.

[40] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Learning feature pyramids for human pose estimation. In *proceedings of the IEEE international conference on computer vision*, pages 1281–1290, 2017.

[41] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[42] Adrian Bulat and Georgios Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *European Conference on Computer Vision*, pages 717–732. Springer, 2016.

[43] Zhiqiang Tang, Xi Peng, Shijie Geng, Lingfei Wu, Shaoting Zhang, and Dimitris Metaxas. Quantized densely connected u-nets for efficient landmark localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 339–354, 2018.

[44] Guanghan Ning, Zhi Zhang, and Zhiquan He. Knowledge-guided deep fractal neural networks for human pose estimation. *IEEE Transactions on Multimedia*, 20(5):1246–1259, 2017.

[45] Diogo C Luvizon, Hedi Tabia, and David Picard. Human pose regression by combining indirect part detection and contextual information. *Computers & Graphics*, 85:15–22, 2019.

[46] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1831–1840, 2017.

[47] Chia-Jung Chou, Jui-Ting Chien, and Hwann-Tzong Chen. Self adversarial training for human pose estimation. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 17–30. IEEE, 2018.

[48] Lipeng Ke, Ming-Ching Chang, Honggang Qi, and Siwei Lyu. Multi-scale structure-aware network for human pose estimation. In *Proceedings of the european conference on computer vision (ECCV)*, pages 713–728, 2018.