Distilling Dynamic Spatial Relation Network for Human Pose Estimation

Kewei Wu[12]
wukewei@hfut.edu.cn
Tao Wang[2]
wanttao2021@163.com
Zhao Xie[12]
xiezhao@hfut.edu.cn
Dan Guo[12]
guodan@hfut.edu.cn

[1] Key Laboratory of Knowledge Engineering with Big Data
Hefei University of Technology, Ministry of Education
Hefei,China

[2] School of Computer Science and Information Engineering
Hefei University of Technology
Hefei,China

Abstract

Human pose estimation is a challenging task that requires the comprehension of the pose structure. This work can refer to spatial relation inference in a pose structure model; how to model the dynamic spatial relation against various unreliable joints is critical. To this end, we propose a Distilling Dynamic Spatial Relation network (DDSR), which builds pose-based graph representation by exploiting the feature of spatial relation from the location distribution of joints. We use a dynamic message propagation mechanism to update the spatial relation on edges. Specifically, to filter out the noisy predictions, we select the joints with high confidence; to enhance the spatial relation in a large receptive field, we propagate multi-stage messages among joints. Besides, to reduce the computation cost of the multi-stage message propagation, we design a cross-resolution distillation framework. We use a new spatial distillation loss to verify the spatial relation between the teacher model and the student model. Experimental results on COCO and MPII datasets show that our method is superior to the state-of-the-art methods. The visualization results further verify the interpretability of our spatial relation.

1 Introduction

Human pose estimation is an important task in the field of computer vision. This paper concentrates on single-person pose estimation, which is the foundation of many applications in multi-person pose estimation [11, 16, 24], video pose estimation [1, 9] and tracking [32].

To learn the joint representation for pose estimation, the CNN-based model [20, 22, 28, 33, 34, 40] describes the relation among joints as a heatmap feature. However, these CNN features do not consider the pose structure of joints, which may lead to prediction errors, especially when the joints encounter are obscured by tanglesome background. In contrast, the graph convolution network (GCN) can handle the relation learning in the structure of pose [25, 30, 36], which propagates messages to generate contextual features from dependable joints. Existing graph models employ the joint structure for various constrained inferences.

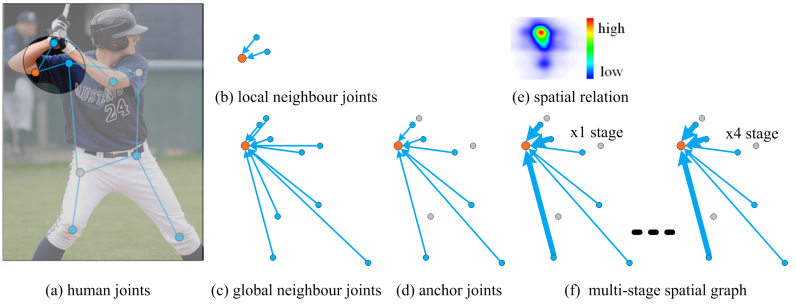


Figure 1: Different relations in the graph structure for pose estimation. In our solution, we focus on exploiting dynamic spatial relations, including anchor joint selection (d) and multi-stage spatial relation inference (f). We learn the relation (edge) weights in the spatial graph (f) of the relative location distribution from the anchor to other joints (e). In (d), blue marks dependable anchor joints, where orange marks the general joint.

The local inference [23, 26, 31, 41], as shown in Fig. 1(b), passes the feature between adjacent joints in the human skeleton, which degrades the messages from the long-term joints. The global inference [2, 3, 8, 10, 18], as shown in Fig. 1(c), considers all the connections between any two joints, which passes all the messages without distinguishing dependable joints. In our view, we should select dependable joints and consider their different (dynamic) relations in the pose structure for contextual representation learning.

As spatial constraint is factual and not be influenced by noisy appearances, we calculate the relative spatial location and embed it into the dynamic propagation to describe the edge weight with a matrix, as shown in Fig. 1(e). In Fig. 1(f), since the graph adopts the spatial distribution-based edge weights, the spatial graph can estimate the deformation of pose variation. For dependable joints selection, we first obtain the heatmap of each joint [28] . Then, we check up the peak value of the heatmap with a threshold to decide whether the joint is dependable. When the message propagates, we update these dependable joints in the graph. Furthermore, to learn more deep relevance of joints in the graph, we adopt the multi-stage graph with multiple times propagations in Fig. 1(f).

However, the multi-stage model introduces vast parameters, especially when the high-resolution backbone model further increases the size of the input feature [5, 28]. As well-known, knowledge distillation [14, 21, 37, 38] is emergent for vast parameters, where the teacher network models multi-stage high-resolution and the student uses the single-stage low-resolution model parameters. In our work, as shown in Fig. 2, the teacher model provides the predicted heatmap as the soft label to train the student model. We adopt the predicted heatmap to design a spatial relation loss to distill the spatial relation of joints in the pose structure.

The contributions are summarized as follows.

(1) We propose a Distilling Dynamic Spatial Relation (DDSR) network, which leverages the spatial location distribution to constraint the pose graph. The edge weight is described by a statistic of relative joint coordinates of pose instances.

(2) We design a dependable dynamic message propagation in the pose graph. We select the dependable joints on the peak value of the joint heatmap. We propagate messages to update the dependable joints with a spatial relation convolution.

(3) To enhance the relation in a large receptive field, we consider a multi-stage network.

To reduce the computational cost, we adopt a cross-resolution distillation scheme. A new spatial distillation loss is designed to verify the spatial relation between the teacher model and the student model. Finally, we obtain a student model with small parameters and promising performance.

(4) Experimental results on COCO and MPII datasets show that DDSR is superior to the state-of-the-art methods. The visualization results further verify the interpretability of our spatial relation.

2 Related Works

Human Pose Estimation. Traditional single-person pose estimation methods used skeleton-based graph structure [25, 30, 36]. Subsequently, a large number of works have been developed with the CNN technique [4, 11, 17, 34, 35, 39, 45]. With the development of CNN, convolutional pose machine [33] and stacked hourglass [22] learned the joint feature by using deep networks. A powerful CNN backbone, High-Resolution Net (HRNet) [28] is widely used [5, 15, 27, 31, 40, 42, 43, 44]. These methods produce high-quality features but ignore the inherent spatial relation of joints. A 2D weight matrix [19, 41] is designed to model spatial relation, but it is a latent variable learning rather than modeling with joint coordinates directly.

Graph Neural Network. Graph neural networks can be divided into two categories. One is applying CNN to graph directly [2, 3, 8, 10, 18] by updating the nodes in the graph. The other is using message passing mechanism [12, 26] in the graph to update both nodes and edges. For the task in the paper, graph neural networks have also been used [6, 23, 30, 31]. We adopt the graph structure with high-confident joints and integrate the spatial message of human structure for pose estimation.

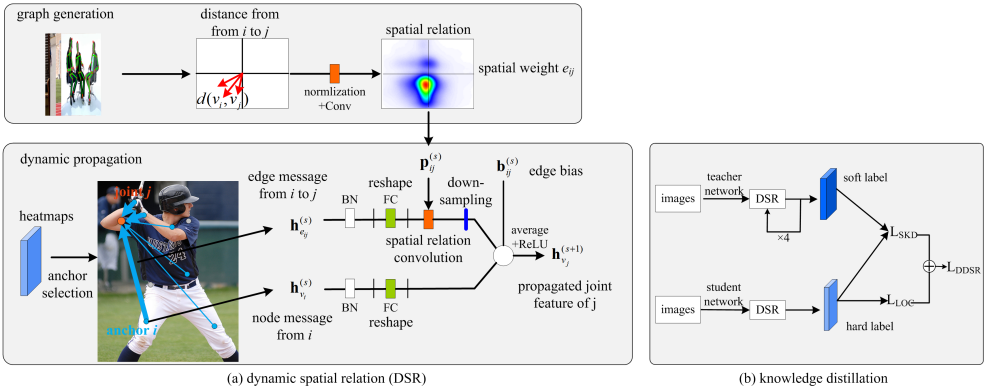
Knowledge Distillation. Knowledge distillation is an effective solution that uses a large trained network to help train a small network [14, 21, 38]. Knowledge distillation can reduce massive parameters in existing pose estimation networks [38]. In the task of pose estimation, knowledge distillation [39] transfers richer structured information of the dense joint confidence graph into a small pose CNNs. Inspired by this, our model makes use of knowledge distillation and designs a new spatial loss function to ensure efficient feature extraction and comprehensive inference of spatial relations.

3 Methodology

In this paper, we propose a human pose detection model named *Distilling Dynamic Spatial Relation (DDSR)* network, which aims at learning spatial relation in the pose graph. The network can be divided into three parts: spatial graph construction referring to anchor joint selection, dynamic message propagation, and spatial knowledge distillation.

3.1 Graph Construction

To build the pose graph, we focus on anchor selection and spatial relation extraction for message propagation. Given the initial joint scores from the backbone, heatmap [28], dependable anchors are selected by a dependable threshold τ, which aims to avoid the message passing from noisy joints. The propagation weights along the edges can be described by the spatial



(a) dynamic spatial relation (DSR) (b) knowledge distillation

Figure 2: The overall framework of *Distilling Dynamic Spatial Relation (DDSR)* network. The spatial graph is propagated based on anchor joint selection and relative location distribution learning (a). Furthermore, the graph is embedded into a knowledge distillation solution with multiple stages (b).

relative location distribution of these dependable anchors. In addition, as shown in Fig. 2, we use multiple stages DSR$\times$ 4 to enlarge the receptive field of the spatial relation.

Anchor Joint Selection. In a skeleton graph of human body $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$, where nodes $\boldsymbol{V} = \{v_i\}$ represent the body joints and edges $\boldsymbol{E} = \{e_{ij}\}$ represent the connection, where i, j are the indexes of the joints. To remove the low-confidence joints in the graph, we realize an anchor selection with a dependable threshold τ. The anchor joints $\boldsymbol{A}$ are decided by the peak value of joint heatmaps as follows:

$$\boldsymbol{A}=\{v_i, peak(\boldsymbol{h}_{v_i}) > \tau\}, \tag{1}$$

where $\boldsymbol{h}_{v_i}$ is the joint heatmap of v_i, i is the joint index, and $peak(\cdot)$ is the peak value function.

Spatial Relation in Graph. Here, we calculate the edge weight to describe the spatial relation (*i.e., relative location distribution*) between the joints. At first, we build a statistical location matrix by counting the distance vector of pairwise joints. As the pairwise joints have directed relations, the distance vectors from joint i to joint j are defined as the location difference of joint j's coordinates substracts joint i's coordinates, *i.e.*, $d(v_i, v_j) = (x_{v_j} - x_{v_i}, y_{v_j} - y_{v_i})$, where (x, y) are the coordinate dimensions. The pairwise joints have the range from $-W$ to W in width and from $-H$ to H in height, where W is the image width and H is the image height. A relative location distribution on all the edges referring to anchors has the shape of $[2W, 2H, N, N]$, where N is the number of all the joints.

To be specific, as shown in Fig, 2.(a), a statistical location matrix $\boldsymbol{D}_{i\to j}$ (the relation from joint i to joint j, $i, j \in [1, N]$) counts the similar distance vectors at each position in the image range of $[2W, 2H]$,

$$\boldsymbol{D}_{i\to j} = \begin{cases} \sum_l \mathbb{I}[(x,y), d^l(v_i, v_j)], & if\ v_i \text{ is an anchor joint;} \\ \boldsymbol{0}, & else \end{cases} \tag{2}$$

where $x \in [-W, W]$, $y \in [-H, H]$. It means that for any position (x, y) in the image area, we collect the homogeneous points (joints) of pose instances. l is the number of pose instances. $\mathbb{I}[\cdot]$ is an indicator function, when $(x, y) = d^l(v_i, v_j)$, it outputs 1, otherwise 0.

Then, based on the relative location matrix $\boldsymbol{D}$, we learn the spatial relation on edge $\boldsymbol{E}$ as $\boldsymbol{P}$, and use a normalization layer and a convolutional layer with a kernel size of 7×7. The operation of each edge is formulated as follows:

$$\boldsymbol{P} = Conv(Norm(\boldsymbol{D})) \in \mathbb{R}^{2W \times 2H}. \tag{3}$$

It is worthy noting that in our graph, here are $\boldsymbol{D}_{i\to j} \neq \boldsymbol{D}_{j\to i}$ and $\boldsymbol{P}_{i\to j} \neq \boldsymbol{P}_{j\to i}$. In other words, the relation between joints i and j has different values along different directions.

3.2 Dynamic Propagation

In the anchor-based joint graph, each node is described with its joint heatmaps. To explore the heatmap difference between two joints, we subtract the heatmap of the joint i from the heatmap of joint j: $\boldsymbol{h}_{e_{ij}} = \boldsymbol{h}_{v_j} - \boldsymbol{h}_{v_i}$. To rich the contextual features of both nodes and edges, we transform the vectorized heatmap and reshape it back to the matrix.

$$\boldsymbol{h}'_{v_i} = \varphi^h(\phi(\varphi^1(BN(\boldsymbol{h}_{v_i})))) \in \mathbb{R}^{W\times H}; \tag{4}$$

$$\boldsymbol{h}'_{e_{ij}} = \varphi^h(\phi(\varphi^1(BN(\boldsymbol{h}_{e_{ij}})))) \in \mathbb{R}^{W\times H}, \tag{5}$$

where BN denotes Batch normalization, $\varphi^1(\cdot)$ denotes the vectorizing function, $\phi(\cdot)$ denotes the FC layer, and $\varphi^h(\cdot)$ is a reshape function.

According to Eq. 5, we know that each node has adaptive neighbor joints according to their respective spatial relation in the image. Thus, we deem that each spatial graph has a dynamic relation obeying the characteristics of the pose instance itself. As the dynamic propagation process shown in Fig. 2 (a), we jointly combine the spatial relation $\boldsymbol{P}_{i\to j}$ and contexts of node $\boldsymbol{h}'_{v_i}$ and edge $\boldsymbol{h}'_{e_{ij}}$ to propagate the message through the edge form i to j as follows:

$$\boldsymbol{h}_{v_j}^{(s+1)} = ReLU\Big(\frac{1}{N^A}\sum_{i\in\boldsymbol{A}}(ds(\boldsymbol{P}_{i\to j}^{(s)} \odot \boldsymbol{h}'^{(s)}_{e_{ij}})) + \boldsymbol{h}'^{(s)}_{v_i} + \boldsymbol{b}_{ij}^{(s)})\Big), \tag{6}$$

where i is an anchor joint, $\boldsymbol{A}$ is the set of selected anchor joints, N^A is the joint number of $\boldsymbol{A}$, and $\boldsymbol{b}_{ij}^{(s)}$ is a bias term. $ReLU(\cdot)$ denotes the Rectified Linear function. The first term in Eq. 6 performs a spatial relation convolution $\odot$ to realize the spatial relation integration (with both relative location distribution and heatmap). ds denotes a downsampling operation with nearest-neighbor interpolation to transform the variable into the shape of $[W,H]$. The second term is to integrate all the heatmaps of anchor neighbors $\{i\} \in \boldsymbol{A}$. At last, s denotes the stage index. We use the above dynamic propagation to compose a multi-stage module to further exploit message propagation deeply.

3.3 Spatial Knowledge Distillation

We design the spatial knowledge distillation scheme for two reasons. On one side, the teacher model provides a soft label of joint distribution which is complementary to a hard label in the form of the ground truth. On the other side, the teacher model can learn high-resolution solutions in the multi-stage mode, which can help the student model with small parameters and less training cost to achieve promising performance.

Generally, Mean Squared Error (MSE) loss is used for single joint location estimation [39]:

$$L_{LOC} = \frac{1}{N}\sum_{i=1}^{N} ||\boldsymbol{m}_{v_i} - \boldsymbol{m}_{v_i}^{gt}||_2^2, \tag{7}$$

where $\boldsymbol{m}_{v_i}$ refers to the confidence heatmap for the joint i, $\boldsymbol{m}_{v_i}^{gt}$ is the ground-truth heatmap, and N is the number of all the joints.

To estimate the spatial relation of edge e_{ij}, a confidence heatmap $\boldsymbol{m}_{e_{ij}}$ is introduced as follows:

$$\boldsymbol{m}_{e_{ij}}(x,y,\sigma) = \frac{1}{2\pi\sigma^2} exp(\frac{-\{[x-(x_j-x_i)]^2-[y-(y_j-y_i)]^2\}}{2\sigma^2}), \tag{8}$$

where (x,y) specifies the joint coordinates and σ denotes a pre-fixed spatial variance hyper-parameter.

The previous distillation function is designed for single-label-based softmax cross-entropy loss [39], which ignores spatial relation among joints. The confidence heatmap is calculated with a Gaussian distribution, which provides a non-linear estimation of the spatial relation. Then, according to the confidence heatmap of each edge, we propose a spatial knowledge distillation (SKD) loss as follows:

$$L_{SKD} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} ||\boldsymbol{m}_{e_{ij}}^S - \boldsymbol{m}_{e_{ij}}^T||_2^2 + \frac{1}{N}\sum_{i=1}^{N} ||\boldsymbol{m}_{v_i}^S - \boldsymbol{m}_{v_i}^T||_2^2 \tag{9}$$

where $\boldsymbol{m}_{e_{ij}}^S$ specifies the confidence heatmap for the edge e_{ij} in the student model, and $\boldsymbol{m}_{e_{ij}}^T$ specifies that in the pre-trained teacher model in the final stage. $\boldsymbol{m}_{v_i}^S$ and $\boldsymbol{m}_{v_i}^T$ are the corresponding confidence heatmaps of joint i obatined from teacher and student models. To distill the spatial knowledge, we formulate the overall loss function as follows:

$$L_{DDSR} = L_{LOC} + L_{SKD}. \tag{10}$$

where L_{LOC} is used to allow the student model learning from the hard label, while L_{SKD} is used to constraint the student model to match the spatial relation heatmap of the teacher model. As the large resolution increases the computational cost in the convolutional layer, we reduce the resolution in the student network. In the model, the joint heatmap and edge obtained from the teacher network have to be downsampled to the same resolution of the student network.

4 Experiment

4.1 Dataset and Implementation

Dataset. COCO. The COCO dataset contains more than 200K images and 250K individual instances, where each instance is marked with 17 key points. We train the model on the COCO train2017 (including 57K images and 150K person instances) and evaluate it on the validation and test-dev set (containing 5K and 20K images, respectively). The evaluation metric uses mAP across the 10 OKS threshold. **MPII.** The MPII Human Pose dataset consists of around 25K images with full-body pose annotations, where there are 12K subjects for

Table 1: Ablation study- anchor selection.

Threshold	Params	GFLOPs	AP	AR
$\tau = 0.30$	32.0M	7.64	73.8	78.8
$\tau = 0.45$	31.3M	7.55	74.4	79.6
$\tau = 0.60$	30.7M	7.49	74.7	80.3
$\tau = 0.75$	30.3M	7.43	**75.1**	**80.4**
$\tau = 0.90$	29.2M	7.27	74.0	79.8

Table 2: Ablation study- multi-stage.

Stage	Params	GFLOPs	AP	AR
2	32.1M	7.73	75.1	80.4
3	34.3M	8.06	75.5	80.6
4	37.8M	8.62	76.1	81.0
5	41.7M	9.22	76.1	**81.2**
6	45.0M	9.71	**76.2**	81.0

testing and the remaining subjects for the training set. We use the PCKh (head-normalized probability of correct keypoint) score as the standard metric.

Training and testing. We upload the pre-trained parameter of HRNet [28] on the COCO dataset to extract the heatmap first and then fine-tune the HRNet part and train the DSR part. The data augmentation follows the HRNet scheme, including the augmentation operations of random scale ($[0.65, 1.35]$), random rotation ($[-45°, 45°]$), flipping, and sampling half body. The anchor threshold τ is an empirical parameter fixed in each stage. We discuss it in the following Sec. 4.2. The spatial convolution $\odot$ is set with Kaiming initialization [13], the kernel parameter 7×7, and is shared in each stage. The spatial variance parameter σ in Eq. 8 for loss estimation is set to 2. At last, the batch size is set to 32. The base learning rate is set to $5e{-}4$, which is dropped to $5e{-}5$ and $5e{-}6$ at the 80th and 160th epochs respectively.

4.2 Ablation Study

To prove the effectiveness of our method, we experiment it over the backbone of HRNet-w32-256×192 [28], and list the results on COCO test-dev set in Tables 1∼5.

Anchor Selection. Table 1 shows the results with various thresholds for anchor selection. It turns out that to avoid faulty message propagation, $\tau = 0.75$ is the best setting. Neither too small nor too large is not appropriate. We choose 0.75 as the anchor selection threshold in the following experiments.

Multi-stage. Table 2 indicates that the performance of our method grows along with the increase of the stage number of DSRs. The performance decreases when the stage is less than 4. And it tends to saturate once it exceeds 4.

Spatial Relation. We test different spatial relations for messages passing along the edges in the graph. As shown in Table 3, "Random" sets the edge weights with random initialization, "Distance" uses the relative distance between joints as edge weight, and "Location distribution" uses the statistical metrics of relative location distribution, *i.e.,* $\boldsymbol{D}$ and $\boldsymbol{P}$ in our graph. The results show that the performance of "Location distribution" increases by 1.2 AP compared with "Distance".

Message Propagation. Here, we compare our dynamic message propagation with the message propagation in local and global perspectives. Table 4 shows that local propagation considers only neighbor joints in the skeleton graph and gets worse performance than global propagation. The global propagation alleviates the noisy messages slightly. When discarding unreliable joints (anchors), our dynamic propagation increases by 1.4 AP compared with the global propagation.

Table 3: Ablation study- spatial relation.

Graph Generation	AP	AP^{50}	AP^{75}	AR
Random	74.5	90.6	82.6	80.4
Distance	74.9	90.7	82.7	80.3
Location distribution	**76.1**	**90.9**	**83.2**	**81.0**

Table 4: Ablation study- propagation.

Propagation	AP	AP^{50}	AP^{75}	AR
Local	74.3	89.8	82.5	80.0
Global	74.7	90.1	82.7	80.3
Dynamic	**76.1**	**90.9**	**83.2**	**81.0**

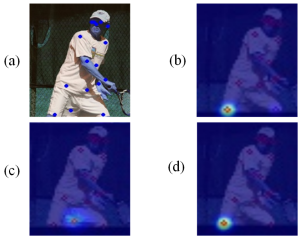


Figure 3: Visualization of hard label (a), soft label (b), HRNet prediction (c), and our DDSR prediction (d) of an pose instance.

Table 5: Evaluating the generality of spatial knowledge distillation (SKD).

SKD	Student Net	AP
×	4-stage Hourglass	66.2
✓	4-stage Hourglass	**69.3**
×	SimpleBaseline-R50	59.3
✓	SimpleBaseline-R50	**62.9**
×	HRNet-w32	66.9
✓	HRNet-w32	**70.3**

Spatial Knowledge Distillation. As our DSR×4 improves the performance while bringing a heavy calculation burden. To decrease computation, we use knowledge distillation with the teacher network DSR×4 with the input size of 384×288 to train the student net with the input size of 128×96. Table 5 shows that, based on different backbone networks, the soft label from the teacher network gives 3.1%, 3.6%, 3.4% AP gain compared with three original networks without the spatial knowledge distillation.

Besides, we visualize the hard label and soft label in Fig.3(a) and Fig.3(b). This indicates that soft label can provide potentially locations that are not provided in the hard label. We find that when the joint exists in the marginal areas of the image, its prediction decoded from the heatmap is obscure and indefinite in Fig.3(c). When we use spatial knowledge distillation, the joint predicted in Fig.3(d) is consistently to the soft label.

4.3 Results on COCO Dataset

In this section, we test the teacher model DSRx4 with the anchor selection threshold of 0.75.

Result on the COCO validation set. As shown in Table 6, the AP of DDSR and DSR×4 increase compared with the backbone of HRNet-w48, robustly on the resolution of 256×192 and 384×288. Results show that the larger resolution boosts up the performance. This indicates that high resolution preserves spatial and structural clues, and benefits accurate spatial relation for message propagation.

Result on the COCO test-dev set. In Table 7, DDSR achieves AP of 76.5, which is 2.8 and 1 higher than SimpleBaseline [34] and HRNet-w48 [28]. Graph-PCNN [31] gets a high performance as it excels at a regression model to relocate the true joint in the candidates. Unlike the Graph-PCNN, our DSR focuses on embedding spatial relations into the GCN operation and modeling the dynamic edge propagation. Furthermore, the teacher model DSRx4 benefits from multi-stage training and high-resolution soft label with the best performance. The teacher model DSR×4 achieves 77.0 AP and 73.6 AP^M, which shows our model detects the poses more effectively.

Table 6: Comparison with HRNet on COCO validation set.

Method	Backbone	Size	Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
HRNet	HRNet-48	256×192	63.6M	14.6	75.1	90.6	82.2	71.5	81.8	80.4
DDSR	HRNet-48	256×192	66.3M	15.8	76.7	91.5	83.7	72.8	82.1	80.7
HRNet	HRNet-48	384×288	63.6M	32.9	76.3	90.8	82.9	72.3	83.4	81.2
DDSR	HRNet-48	384×288	67.5M	34.7	77.5	92.5	84.5	73.3	83.5	81.3
DSR×4	HRNet-48	384×288	79.2M	40.1	**78.2**	**93.1**	**85.3**	**74.4**	**84.1**	**81.4**

Table 7: Comparison with the state-of-the-art methods on COCO test-dev set.

Method	Backbone	Size	Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Macro–Micro [45]	Hourglass	256×192	27.1M	23.5	73.7	91.9	81.7	70.6	79.3	79.1
RMPE [11]	PyraNet	256×256	45.0M	11.0	67.8	88.2	74.8	63.9	74.0	-
HRNet [28]	HRNet-w32	384×288	28.5M	16.0	74.9	92.5	82.8	71.3	80.9	80.1
DDSR	HRNet-w32	384×288	30.3M	17.2	75.5	92.5	83.0	72.4	81.9	81.0
HRNet [28]	HRNet-w48	384×288	63.6M	32.9	75.5	92.5	83.3	71.9	81.5	80.5
DARK [40]	HRNet-w48	384×288	63.6M	32.9	76.2	92.5	83.6	72.5	82.4	81.1
UDP [15]	HRNet-w48	384×288	63.8M	33.0	76.1	92.5	83.5	72.8	82	81.3
Simple Baseline [34]	ResNet-152	384×288	68.6M	35.6	73.7	91.9	81.1	70.3	80.0	79.0
TNet-D3W96 [43]	D3W96	384×288	-	-	75.8	92.6	83.6	72.7	81.4	81.1
G-PCNN [31]	HRNet-w48	384×288	-	-	76.8	92.6	84.3	73.3	**82.7**	**81.6**
DDSR	HRNet-w48	384×288	67.5M	34.7	76.5	93.3	83.9	73.3	82.2	81.5
DSR×4	HRNet-w48	384×288	79.2M	40.1	**77.0**	**93.8**	**84.4**	**73.6**	82.5	**81.6**

4.4 Results on the MPII test set.

Table 8 shows that DDSR achieves 92.5 PKCh@0.5, and outperforms the stacked hourglass [22] and HRNet [28]. The teacher model DSR×4 increases PKCh@0.5 by 0.4 compared with HRNet-w32, and the student model DDSR still has a slight increase. The reason might be that the MPII dataset has simple pose graphs, the joint connections are easier to learn with only appearance feature.

Table 8: Comparisons with the state-of-the-art methods on MPII test set (PCKh@0.5).

Method	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Total
CPM [33]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Hourglass[22]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Simple Baseline [34]	98.5	96.6	91.9	87.6	91.1	88.1	84.1	91.5
GAN-pose[7]	98.2	96.8	92.2	88.0	91.3	89.1	84.9	91.8
Adversarial PoseNet [2]	98.1	96.5	92.5	88.5	90.2	89.6	86.0	91.9
Structure-aware Network [17]	98.5	96.8	92.7	88.4	90.6	89.3	86.3	92.1
Compositional Model [29]	98.4	96.9	92.6	88.7	**91.8**	89.4	86.2	92.3
HRNet-w32 [28]	98.6	96.9	92.8	89.0	91.5	89.0	85.7	92.3
Macro–Micro [45]	98.5	**97.0**	92.8	88.9	91.5	**89.9**	86.4	92.5
DDSR	**98.8**	96.8	92.7	89.3	91.6	89.3	86.2	92.5
DSR×4	**98.8**	96.8	**93.0**	**89.6**	91.7	89.5	**86.6**	**92.7**

5 Conclusion

In this paper, we propose a Distilling Dynamic Spatial Relation network (DDSR) for single-person pose estimation, which models spatial relation learning based on dynamic message passing and knowledge distillation. The proposed DDSR achieves state-of-the-art performance on both COCO dataset and MPII dataset. The visualization results further show that the pose prediction is more accurate with effective spatial relation inference.

6 Acknowledgments

This research was supported by the National Key Research and Development Program of China (2017YFB1002203) and the National Nature Science Foundation of China (61876058).

References

[1] Bruno Artacho and Andreas Savakis. Unipose: Unified human pose estimation in single images and videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7035–7044, 2020.

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

[3] Jie Chen, Tengfei Ma, and Cao Xiao. Fastgcn: fast learning with graph convolutional networks via importance sampling. *arXiv preprint arXiv:1801.10247*, 2018.

[4] Yu Chen, Chunhua Shen, Xiu-Shen Wei, Lingqiao Liu, and Jian Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1212–1221, 2017.

[5] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5386–5395, 2020.

[6] Yu Cheng, Bo Wang, Bo Yang, and Robby T. Tan. Graph and temporal convolutional networks for 3d multi-person pose estimation in monocular videos. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, pages 1157–1165. AAAI Press, 2021.

[7] 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.

[8] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *arXiv preprint arXiv:1606.09375*, 2016.

[9] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation from multiple views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7792–7801, 2019.

[10] David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. *arXiv preprint arXiv:1509.09292*, 2015.

[11] 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.

[12] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pages 1263–1272. PMLR, 2017.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1026–1034. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.123. URL https://doi.org/10.1109/ICCV.2015.123.

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[15] Junjie Huang, Zheng Zhu, Feng Guo, and Guan Huang. The devil is in the details: Delving into unbiased data processing for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5700–5709, 2020.

[16] Sheng Jin, Wentao Liu, Wanli Ouyang, and Chen Qian. Multi-person articulated tracking with spatial and temporal embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5664–5673, 2019.

[17] 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.

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

[19] Deying Kong, Haoyu Ma, and Xiaohui Xie. SIA-GCN: A spatial information aware graph neural network with 2d convolutions for hand pose estimation. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020.

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[21] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5191–5198, 2020.

[22] 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.

[23] Xuecheng Nie, Jiashi Feng, Junliang Xing, and Shuicheng Yan. Pose partition networks for multi-person pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 684–699, 2018.

[24] 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.

[25] Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Poselet conditioned pictorial structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595, 2013.

[26] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 401–417, 2018.

[27] 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.

[28] 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.

[29] Wei Tang, Pei Yu, and Ying Wu. Deeply learned compositional models for human pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 190–206, 2018.

[30] Jonathan Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. *arXiv preprint arXiv:1406.2984*, 2014.

[31] 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.

[32] Manchen Wang, Joseph Tighe, and Davide Modolo. Combining detection and tracking for human pose estimation in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11088–11096, 2020.

[33] 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.

[34] 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.

[35] 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.

[36] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR 2011*, pages 1385–1392. IEEE, 2011.

[37] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017.

[38] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisit knowledge distillation: a teacher-free framework. *arXiv preprint arXiv:1909.11723*, 2019.

[39] Feng Zhang, Xiatian Zhu, and Mao Ye. Fast human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3517–3526, 2019.

[40] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7093–7102, 2020.

[41] Hong Zhang, Hao Ouyang, Shu Liu, Xiaojuan Qi, Xiaoyong Shen, Ruigang Yang, and Jiaya Jia. Human pose estimation with spatial contextual information. *arXiv preprint arXiv:1901.01760*, 2019.

[42] Jiabin Zhang, Zheng Zhu, Jiwen Lu, Junjie Huang, Guan Huang, and Jie Zhou. Simple: Single-network with mimicking and point learning for bottom-up human pose estimation. *arXiv preprint arXiv:2104.02486*, 2021.

[43] Lei Zhao, Jun Wen, Pengfei Wang, and Nenggan Zheng. Context-guided adaptive network for efficient human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3492–3499, 2021.

[44] Lu Zhou, Yingying Chen, Yunze Gao, Jinqiao Wang, and Hanqing Lu. Occlusion-aware siamese network for human pose estimation. In *European Conference on Computer Vision*, pages 396–412. Springer, 2020.

[45] Lu Zhou, Yingying Chen, Congqi Cao, Yakui Chu, Jinqiao Wang, and Hanqing Lu. Macro-micro mutual learning inside compositional model for human pose estimation. *Neurocomputing*, 449: 176–188, 2021.