Structured Spatial Reasoning for Human Pose Estimation

Ying Huang\(^1\)  
yw52@hznu.edu.cn  
Shanfeng Hu\(^2\)  
shanfeng2.hu@northumbria.ac.uk  
Zi-Ke Zhang\(^3\)  
zkz@zju.edu.cn

\(^1\) Hangzhou Normal University  
Hangzhou, China  
\(^2\) Computer and Information Sciences  
Department  
Northumbria University  
Newcastle upon Tyne, UK  
\(^3\) Zhejiang University  
Hangzhou, China

Abstract

Human pose estimation from single images has made significant progress in the past but still faces fundamental challenges from the occlusion and overlapping of joints in many cases. This is partly due to the limitation of the traditional paradigm for this problem, which attempts to locate human body joints solely and as a result can fail to resolve the spatial connections among joints that are critical for the identification of the whole pose. To overcome this shortcoming, we propose to explicitly incorporate spatial reasoning into pose estimation by formulating it as a structured graph learning problem, in which each image pixel is a candidate graph node with every two nodes connected via an edge that captures their affinity. The advantage of this representation is that it allows us to learn feature embeddings for both the nodes and edges, thereby providing a sufficient capacity to delineate correct human body joints and their connecting bones. To facilitate efficient learning and inference, we exploit self-attention transformer architectures that fuse node and edge learning pathways, which can save parameter numbers and permit fast computation. Experiments on the popular MS-COCO Human pose estimation benchmark show that our method outperforms representative methods.

1 Introduction

Powered by advances in machine learning and deep learning, computer vision applications have made significant progress in recent years, among which human pose estimation is a rapidly evolving one that impacts several human-centered technologies in 3D space, such as virtual reality [4], smart home [20], human-computer interaction [34], and urban brain [37]. The aim of human pose estimation is to identify each joint position of the human body from a given image to obtain the geometric and spatial configuration of the body. This is a fundamentally challenging task because the variations of image appearance and body configuration can be unlimited, which requires powerful spatial reasoning to resolve ambiguous cases when certain joints are occluded or overlapped.
Despite the inherently coherent structure of human body joints, the current mainstream approach to human pose estimation remains largely oblivious of the structure and operates in an object classification fashion for each individual joint [17, 30]. One typical working assumption of this approach is that there can be at most one human joint of the same semantic category at each image region of interest. This may work well in simple cases where all joints are clearly delineated in an image but can fail when some of these joints are not visible due to occlusion or overlapping with others. Take Figure 1 for example. When certain body parts are overlapped or share a similar appearance with other parts, current methods can be misled to generate incorrect predictions using only separate node information.

While spatial reasoning is critical for robust pose identification in challenging cases as illustrated in Figure 1, currently there are few relevant studies to tackle this problem satisfactorily. The method of [14] trains a graph neural network to predict the edges between pairs of selected nodes, but by assuming the existence of only one type of body joint within a single image region, it cannot learn graph node embeddings that distinguish overlapped joints. In contrast, the method of [24] learns joint embeddings as scalars from the spatial locations of joints, which are then used to determine the affinity between joints according to the intra-class distances of these embeddings. Still, the method is not sufficiently flexible to resolve overlapping cases.

In view of the limitations of the existing methods, we propose to explicitly incorporate spatial reasoning into human pose estimation by not just learning to recognize the appearance of individual body joints but also capturing their mutual connections relating to the structure of the whole body posture. Due to the flexibility of graphs for structural representation [31], we propose to formulate pose identification as a structured graph learning problem, in which each image pixel is viewed as a graph node that encodes the visual and spatial features of a potential body joint. Each pair of nodes is connected by an edge that corresponds to a potential bone segment, which provides additional rigidity constraints on the spatial configuration of body joints. Compared with the work of [14] that only captures the spatial
connectivity among a handful of nodes for pose inference, our representation is considerably
more flexible for encoding the affinity between every pair of potential nodes on the image
space, which provides a sufficient capacity to reason the plausible spatial configuration of
potentially challenging body poses.

The core contribution we make in this paper is that we propose a structured graph learn-
ing model on the self-attention transformer architecture that can efficiently perform flexi-
ble spatial reasoning for human pose estimation. While establishing an explicit connection
between every pair of potential joint positions is computationally infeasible, we achieve si-
multaneous node and edge prediction by learning shareable pixel-wise node embeddings
via a popular self-attention model [18] from raw image features and incorporating every
pair of node features to calculate their edge strength. The prediction of joint positions is
done by learning category-specific token embeddings to query each node’s features to pro-
duce human-understandable heatmaps. To enable joint training, we deploy two loss func-
tions for node and edge prediction respectively. The first loss function is the mean squared
error (MSE), which calculates the distance between the predicted joint heatmaps and the
groundtruth annotations. The other loss function we use for edge prediction is the binary
cross-entropy loss that penalizes incorrect prediction of edge presence.

We conduct standard performance benchmarking on the widely used MS-COCO 2018
keypoint detection dataset [20]. Compared with other state-of-the-art methods, including
SimpleBaseline [31], HRNet [26], TransPose [18], and TokenPose [18], our approach achieves
a record-high 77.7 AP on the COCO validation set [20]. Our accuracy improves over that
achieved by the TokenPose [18] method by 1.9 points under the same experimental settings
with a similar level of parameter numbers and computation cost. This is particularly signifi-
cant in that TokenPose only leverages joint prediction for pose estimation, which shows that
our approach of graph-based spatial reasoning is able to provide an additional performance
improvement while not incurring extra computational burdens.

The structure of the paper is as follows. We survey the existing methods for human pose
estimation in Section 2, through which we point out the unique advantage of our method
in data-driven spatial reasoning. We then describe our graph learning approach in detail
in Section 3 and present the experimental results in Section 4. The paper is concluded in
Section 5 with a discussion on future work.

2 Related Work

Human pose estimation has been an active research area for many years. The task of pose
estimation is two-fold. Given an image, one is to locate the position of each body joint and
identify the semantic category it belongs to, and the other is to parse the detected joints and
distinguish those belonging to oneself and other people [41]. The challenge of the task stems
from the fact that there can be unlimited complexity and variability of human posture, cloth-
ing appearance, and background clutters in an image. On top of this are frequent occlusions
and overlapping, which require effective spatial reasoning to resolve the ambiguity in many
cases [11]. In the following, we discuss recent methods that leverage deep learning (con-
volutional neural networks [26], transformers [36]) for this challenge as they are often top
performers on popular evaluation benchmarks [20].

For single-person pose estimation, the challenge is somehow alleviated as the working
assumption is that there is only one person in a given image and therefore there is no need
to determine which person an articulated limb belongs to because only one of each type
of articulated limbs exists in the image [5]. Many solutions are proposed to improve the estimation accuracy, including global information collection [30], multi-scale learning [16], graph representation [27], recursive attention model [6], spatial correlation [38], etc. The evaluation of these methods on the MPII single-person pose estimation dataset shows that correlation relationship modelling has a key role in the human pose estimation task.

For multi-person pose estimation, the difficulty is lifted compared with that in the single-person case since the process of joint parsing is required to determine the correct belonging of each detected joint to each person [1]. Currently, there are two main lines of work that tackle this challenge. The top-down approach works by detecting each human region first and then performing single-person pose estimation in each region separately, hence removing the need of attaching the detected joint limbs to the correct body. The proposed approaches include pose spatial transformation [2], cascaded pyramid network [8], channel and spatial information enhancement [25], constructing a graph network [29], attention model [35]. The current top-down methods have high accuracy in pose estimation in simple scenes, but when there is a severe occlusion in the scene, such methods are more difficult to utilize partially visible joint cues and the algorithm accuracy will be significantly degraded. Our approach introduces graph learning and strengthen the network supervision by both node and edge information. The bottom-up approach, compared with the top-down one, is a local-to-global process that first detects all body joints from an image and then divides them and attributes them to each target individual person. The interesting studies include high-resolution network HigherHRNet [4], explicit joint-limb association [3], joint-limb complementarity [1, 2], graph network-based joint clustering [14], joint attribution grouping [24]. Compared with our approach, those methods do not handle the cases of overlapped or occluded joints explicitly.

As our proposed method exploits graphs for human pose spatial reasoning, here we also discuss methods that explicitly address graph learning. The graph is an effective representation for structured and connected data. Graph learning refers to the use of machine learning methods on graphs to obtain relevant features. Here we mainly review the methods to study the correlation between two graph nodes. The related work includes link prediction [19, 21], structural predictability [15], robustness analysis of similarity metrics [40], drug-target interaction prediction [22], evolutionary neural network-based model [33], local subgraph representation [1], graph attention networks [10, 28]. It is worth pointing out that [14] learn to determine the connectivity between human joints by graph neural networks, which is similar to the concept of link prediction, but the method still assumes that there is only one joint of the same type at the same location. Currently, graph learning has not been deeply investigated yet in the field of human pose estimation. Considering that human pose is a type of structured graph data, and the difficulties in practical applications, introducing graph learning can provide more space for technology exploration.

3 Structured Spatial Reasoning for Pose Estimation

Human pose data can be viewed as structured graph data, which contains information about human nodes and connected edges. At the graph level, nodes and edges are two necessary elements for a graph representation, and similarly, human body nodes and connecting edges are critical constituents for the human body pose. In this paper, graph learning refers to learning feature embeddings of image information, and then predicting human joints and connected edges in the image.
Figure 2: Overview of our graph learning architecture for human pose estimation. The convolutional feature maps are uniformly divided into patches and linearly transformed to 1D visual tokens. 1D human joint category vectors are randomly initialized as query tokens. Then, visual and query tokens are combined as input into the Transformer encoder to learn multi-modal information through the multi-head self-attention algorithm. The last outputs of visual tokens are spliced and reshaped back to the size of CNN feature maps, and pixel-level node embeddings are obtained. Finally, category and node embeddings are used to predict human joints and edges, respectively.
For the task of outputting human joint information in images, many deep learning-based methods have been studied before, such as using deep convolutional neural networks, transformer attention models, etc. To simplify the model complexity, we fuse two tasks as multi-task learning using a single model, instead of designing two separate models. This requires us to make full use of model potentialities.

The useful human pose information usually includes the joint category, location, appearance, edge orientation, and contextual background. The idea of existing graph learning algorithms in determining whether an edge exists between two nodes is to calculate the similarity of two nodes’ embeddings. The higher the similarity, the more likely an edge to exist. Simultaneously, node recognition also relies on its feature embedding, which provides a common point for both tasks of graph learning.

From this point, we design a model with a parameter-shared main stem and two lightweight task-oriented heads. The main stem of the model is used to learn spatial feature embeddings, then two heads perform separate tasks. Specifically, for the human joint prediction task, the model outputs the predicted joint heatmaps. For the edge prediction task, the feature embeddings of two joints and the corresponding two joint category embeddings are concatenated and then input into a multilayer perceptron. The model architecture is illustrated in Figure 2. The technical details of implementing these two specific tasks are described below.

### 3.1 Node Prediction

Currently, the popular network framework is first uses deep convolutional neural network as the base network backbone to extract the high-level features of the image, and then use the Transformer attention model to learn the global association relationship from the feature maps [12, 33]. At last, the framework predicts \(N\) heatmaps of size \(\hat{H} \times \hat{W}\), each heatmap corresponds to a human joint category, and the location with a peak on the heatmap is considered to be the position where that type of joint is located. This combined framework achieves state-of-the-art performance with less than half of the amount of parameters. Considering that this framework is more flexible, our paper also follows this framework. Given an input image \(I \in \mathcal{R}^{H \times W \times 3}\), the backbone network model extracts convolutional features from the image, and then generates feature maps \(F \in \mathcal{R}^{\hat{H} \times \hat{W} \times C}\). To fit the dimensionality of the attention model, \(F\) is uniformly divided into \(L = \frac{\hat{H}}{P_h} \times \frac{\hat{W}}{P_w}\) patches of size \(P_h \times P_w\). These blocks are further reshaped to 1D vectors of size \(P_h \times P_w \times C\), and each vector is adapted to a \(d\)-dimensional embedding \(v \in \mathcal{R}^d\) via linear transform \(P \rightarrow v\). Since human pose estimation needs to output the position of human joints, the two-dimensional position embedding \(p_i\) is also added to every vectors to generate visual tokens \(v'_i = v_i + p_i\), where \(i \in [1, L]\). The transformed visual tokens represent the embeddings of spatial regions.

According to the attention mechanism, we first pre-define \(N\) learnable \(d\)-dimensional vectors as category query tokens, which represent \(N\) human joint categories respectively. Once the query tokens and visual tokens are obtained, they enter into the commonly used multiple attention modules and learn new embeddings according to a general attention formulation:

\[
f'_i = f_i + \sum_j \text{softmax}_j \left( s \left( f'_i, f^K_j \right) \right) T_V \left( f^K_j \right)
\]

where \(s\) denotes the similarity function of the \(i\)-th query instance feature \(f'_i\) and the \(j\)-th key instance feature \(f^K_j\), and \(T_V(\cdot)\) is the linear transformation of \(j\)-th value instance feature.
After multiple stacked multi-head attention modules, the final computed category embeddings are mapped into $\hat{H} \times \hat{W}$-dimensional feature vectors via linear projection, and then turned into $N$ two-dimensional heatmaps by reshape operation. The MSE loss function is used to calculate the difference between the groundtruth and predicted heatmaps to get the loss $l_{kpt}$ during training.

### 3.2 Edge Prediction

Edge prediction means that for a set of nodes $V$, predicting the set $E$ of observable edges from a series of node combinations [39], and more specifically, for the adjacency matrix $A$ composed of node sets, $A_{ij}=1$ if the node pair $(i, j)$ belongs to $E$, otherwise $A_{ij}=0$. To simplify the calculation, we take the heatmaps generated in the node prediction task as a reference, and consider each pixel of heatmaps as a node, the total number of nodes is $\hat{H} \times \hat{W} \times N$. Then we extract feature embeddings for these nodes. In the node prediction task, we have obtained visual tokens with position embedding for every image patch. We can reconvert and combine these visual tokens back to the size $\hat{H} \times \hat{W} \times d$, results in $d$-dimensional embeddings for every spatial position. To overcome the case of overlapped nodes, the embedding of a node and its corresponding category embedding are concatenated. Such two extended embeddings of an edge are input to a multilayer perceptron, and finally we can obtain the prediction probability of this edge. The design of combining category and visual embeddings can enhance the representation of an edge, fusing the information including location, category, appearance, and global relationship. The binarized cross-entropy loss function is used for training. Positive edge samples are obtained from groundtruth, while the rest of candidate edges are as negative edge samples. The predicted probability values of these positive and negative samples are compared with the true label values 0/1 to obtain the loss $l_{pair}$.

Node prediction and connected edge prediction are combined as joint training:

$$loss = l_{kpt} + \lambda \times l_{pair}$$

(2)

where $\lambda$ is a balance weight. Note that in edge prediction, node features are converted from visual tokens and no new parameters added, the only added model parameter is the lightweight three-layer perceptron used in edge prediction. In the experimental section, we will see that joint learning can improve the performance of human pose estimation.

### 4 Experimental Details

#### 4.1 MS-COCO Keypoint Dataset

The qualitative and quantitative experiments are performed on the MS-COCO 2018 keypoint detection dataset [20]. This popular dataset contains training, validation and testing sets. On the training and validation sets, there are 118,287 and 5000 images respectively, a total of over 150,000 human instances with around 1.7 million labelled keypoints. The testing set has two splits: test-dev and test-challenge, each includes roughly 20,000 images. We train and evaluate our models on the training and validation sets. The model is also evaluated on the test-dev set and accuracy values are obtained from the online evaluation server.

In order to match predictions to groundtruth, COCO keypoint dataset defined an overlapping index suitable for human pose data, object keypoint similarity (OKS), which calculates the overlapping ratio between groundtruth and predictions in terms of point distribution.
Table 1: Comparisons of our model to other state-of-the-art models. Hybrid convolution plus Transformer methods (TokenPose[18], TransPose[35], and ours) outperform pure convolution based methods (SimpleBaseline[32] and HRNet[26]). "+" means model ensemble. Particularly, with similar model parameters and settings, AP of our approach is higher than TokenPose by 2.7 points.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Input size</th>
<th>#Params</th>
<th>GFLOPs</th>
<th>AP</th>
<th>AP^{50}</th>
<th>AP^{75}</th>
<th>AP^{M}</th>
<th>AP^{L}</th>
<th>AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple-Res50</td>
<td>256×192</td>
<td>34.0M</td>
<td>8.9</td>
<td>70.4</td>
<td>88.6</td>
<td>78.3</td>
<td>67.1</td>
<td>77.2</td>
<td>76.3</td>
</tr>
<tr>
<td>Simple-Res101</td>
<td>256×192</td>
<td>53.0M</td>
<td>12.4</td>
<td>71.4</td>
<td>89.3</td>
<td>79.3</td>
<td>68.1</td>
<td>78.1</td>
<td>77.1</td>
</tr>
<tr>
<td>Simple-Res152</td>
<td>256×192</td>
<td>68.6M</td>
<td>15.7</td>
<td>72.0</td>
<td>89.3</td>
<td>79.8</td>
<td>68.7</td>
<td>78.9</td>
<td>77.8</td>
</tr>
<tr>
<td>HRNet-W32</td>
<td>256×192</td>
<td>28.5M</td>
<td>7.1</td>
<td>74.4</td>
<td>90.5</td>
<td>81.9</td>
<td>70.8</td>
<td>81.0</td>
<td>79.8</td>
</tr>
<tr>
<td>HRNet-W48</td>
<td>256×192</td>
<td>63.6M</td>
<td>14.6</td>
<td>75.1</td>
<td>90.6</td>
<td>82.2</td>
<td>71.5</td>
<td>81.8</td>
<td><strong>80.4</strong></td>
</tr>
<tr>
<td>PureTransformer</td>
<td>256×192</td>
<td>5.8M</td>
<td>1.3</td>
<td>65.6</td>
<td>86.4</td>
<td>73.0</td>
<td>63.1</td>
<td>71.5</td>
<td>72.1</td>
</tr>
<tr>
<td>TokenPose-L/D24</td>
<td>256×192</td>
<td>27.46M</td>
<td>10.98</td>
<td>75.0</td>
<td>89.7</td>
<td>81.9</td>
<td>71.7</td>
<td>81.8</td>
<td>80.3</td>
</tr>
<tr>
<td>TransPose-H-A6</td>
<td>256×192</td>
<td>17.5M</td>
<td>21.8</td>
<td>75.0</td>
<td>89.8</td>
<td>81.9</td>
<td>71.7</td>
<td>81.7</td>
<td>80.2</td>
</tr>
<tr>
<td>Ours</td>
<td>256×192</td>
<td>27.47M</td>
<td>10.99</td>
<td>75.3</td>
<td>90.6</td>
<td>82.6</td>
<td>72.3</td>
<td>79.5</td>
<td>80.1</td>
</tr>
<tr>
<td>Ours +</td>
<td>256×192</td>
<td>27.47M</td>
<td>10.99</td>
<td><strong>77.8</strong></td>
<td><strong>93.6</strong></td>
<td><strong>84.8</strong></td>
<td><strong>74.9</strong></td>
<td><strong>81.9</strong></td>
<td>80.2</td>
</tr>
</tbody>
</table>

Based on the OKS index, we use six evaluation metrics by adjusting the thresholds of matching criteria to compare the performance of a model. They are AP (i.e. average precision), AP^{50}, AP^{75}, AP^{M}, AP^{L} and AR (i.e. average recall). The 20 top-scoring predictions are selected to attend the evaluations per image.

4.2 Training Details

In this paper, the two-stage top-down human pose estimation paradigm is adopted. In this paradigm, human regions are firstly obtained by a person detector. Then each human instance after cropping and scaling is input to our model and keypoints and edges are predicted. To facilitate the comparisons, we follow the previous methods[18, 35] to use person detectors provided by SimpleBaseline[32]. The CNN backbone of our method is selected from HRNet-W48[26] and its parameters are also initialized by the pre-trained model of HRNet-W48. The Transformer parts of our model are trained from scratch. During training, the Adam optimizer is utilized with mini-batches of size 16. The learning rate is started from 1e-3, and is declined to 1e-4 and 1e-5 at the 200th and 260th epochs, respectively. The total training process iterates 300 epochs on the training set. For the edge sampling, the annotated human joints can form the human skeleton, and skeletons are used as human edges. In the calculation of edge loss, the weight of negative to positive edges is maintained as 5. The random data augmentation is analogous to the steps in HRNet[26].

4.3 Comparison Results

The experimental results and statistics of ours and other methods on the validation set are recorded in Table 1. Hybrid convolution plus Transformer methods (TokenPose[18], TransPose [35], and ours) outperform pure convolution based methods (SimpleBaseline[32] and HRNet[26]) using much fewer parameters, which shows the advantage of Transformer [8] in model parameter reduction. More importantly, previous methods only attempt to locate body joints solely. Compared to TokenPose [18] with similar model parameters and settings, our approach has improved AP by 2.7 points. Compared with TransPose [35] which use a pixel-wise token embedding, our model has less computation cost. We show in Table 2 that our method without the MLP layer performs on par with TokenPose. Considering the key
metric $AP^{50}$ already achieves greater than 90, the further improvement of 0.3 brought by the addition of this design is actually large, especially in the case of a similar amount of parameters and FLOPs. The examples of predicted accurate human joints and heatmaps are shown in Figure 3, which contains some abnormal cases, such as scale, appearance and viewpoint variation. The algorithm can locate human bodies accurately. The qualitative results in Figure 4 validate the superiority of our method in handling heavy occlusion and overlapping. For these cases, structured spatial reasoning provides stronger feature representation. We compare with two additional graph-based methods in Table 3 and show improvement.

In extensive experiments, we found some interesting phenomena. The first point is that negative edge sampling should be fixed during training. We have tried several sampling strategies, including random and sparse sampling. However, these sampling methods can cause undesirable divergent behaviours, and the trained model lacks prediction generalizability, even injecting spatial position encoding for every sampling location. This point validates that a holistic structure should be maintained in graph learning. The second point is extended from the previous problem. For fixed graph sampling, there is a trade-off between the interval of sampling on the image and computational costs. Given sufficient time this line of research will provide further detail.

## 5 Conclusion and Future Work

We have introduced a new approach to explicitly incorporate spatial reasoning into human pose estimation to improve detection accuracy. The key advantage of our method over existing ones is its ability to capture the global pairwise connection among potential joint nodes on the image space, which provides a sufficient capacity to resolve challenging body postures. While encoding the full spatial relationship as a graph is computationally infeasible, we have achieved efficient reasoning by learning shareable pixel-wise node embeddings that can be used to make edge predictions via a jointly trained model. Experimental results show a considerable accuracy improvement over the current state-of-the-art methods on the challenging MS-COCO benchmark. In the future, we plan to use our method in several downstream tasks such as continuous pose tracking in videos.
Figure 3: Examples of predicted human joints and heatmaps.

Figure 4: Qualitative results of our method.
References


