RSGNet: Relation based Skeleton Graph Network for Crowded Scenes Pose Estimation

Yan Dai,[1]* **Xuanhan Wang,**[1] **Lianli Gao,**[1] **Jingkuan Song,**[1 2] **Heng Tao Shen**[1]

[1]Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, 611731

[2]Key Laboratory of Artificial Intelligence, Ministry of Education, Shanghai, 200240

yandai1019@gmail.com, wxuanhan@hotmail.com, lianli.gao@uestc.edu.cn,
jingkuan.song@gmail.com, shenhengtao@hotmail.com

Abstract

Despite of the recent great progress on multi-person pose estimation, existing solutions still remain challenging under the condition of "crowded scenes", where RGB images capture complex real-world scenes with highly-overlapped people, severe occlusions and diverse postures. In this work, we focus on two main problems: 1) how to design an effective pipeline for crowded scenes pose estimation; and 2) how to equip this pipeline with the ability of relation modeling for interference resolving. To tackle these problems, we propose a new pipeline named ***Relation based Skeleton Graph Network (RSGNet)***. Unlike existing works that directly predict joints-of-target by labeling joints-of-interference as false positive, we first encourage all joints to be predicted. And then, a **Target-aware Relation Parser (TRP)** is designed to model the relation over all predicted joints, resulting in a target-aware encoding. This new pipeline will largely relieve the confusion of the joints estimation model when seeing identical joints with totally distinct labels (e.g., the identical hand exists in two bounding boxes). Furthermore, we introduce a **Skeleton Graph Machine (SGM)** to model the skeleton-based commonsense knowledge, aiming to estimate the target pose with the constraint of human body structure. Such skeleton-based constraint can help to deal with the challenges in crowded scenes from a reasoning perspective. Solid experiments on pose estimation benchmarks demonstrate that our method outperforms existing state-of-the-art methods. The code and pre-trained models are publicly available online[1].

Introduction

2D multi-person pose estimation has been a fundamental problem in computer vision, which aims to detect human anatomical joints (e.g., neck, wrist) from a given image. It has attracted huge interest, since it supports wide applications, such as human parsing (Fang et al. 2018; Wang et al. 2018), human-computer interaction (Shotton et al. 2013; Gao et al. 2020), and tracking (Xiao, Wu, and Wei 2018; Song et al. 2018).

Benefiting from the success of deep convolution neural networks (CNNs) and released large-scale datasets such as MSCOCO (Lin et al. 2014), recent proposed pose estimation methods have achieved great progress. They can be roughly divided into *bottom-up* methods and *top-down* methods.

Bottom-up Methods. In general, bottom-up methods firstly detect all human joints, and then group them into different person instances. Although these different methods vary in network topology, most of them (Newell, Huang, and Deng 2017; Cao et al. 2018; Papandreou et al. 2018; Jin et al. 2020; Nie et al. 2019) focus on one problem: how to group candidate joints into individual person instance. For example, (Cao et al. 2018) propose the Part Affinity Field (PAF) to associate joints through their affinity scores, while (Newell, Huang, and Deng 2017) introduce the Associative Embedding that assigns different person tags to joints for grouping. Different from these methods, (Cheng et al. 2020) focus on the quality of candidate joints generation, and propose a scale-aware representation learning method that generates candidate joints from high resolution pose embedding.

Top-down Methods. Unlike bottom-up methods, top-down methods firstly detect out different person instances, and then solve the single person pose estimation problem among regions. Most current state-of-the-art top-down methods (Sun et al. 2019; Xiao, Wu, and Wei 2018; He et al. 2017; Chen et al. 2018; Fang et al. 2017; Li et al. 2019; Qiu et al. 2020; Golda et al. 2019; Guo et al. 2019; Wang et al. 2020b) focus on the second step, that is, how to design a robust neural network for single pose estimation among cropped region images. For example, (Chen et al. 2018) propose to estimate person joints in an easy-to-hard order. Moreover, (Xiao, Wu, and Wei 2018) demonstrate that a simple convolution neural network (e.g., ResNet-50) with a couple of deconvolutions can achieve encouraging pose estimation results. Recently, HRNet is proposed to estimate pose from rich object details by preserving high feature resolution (Sun et al. 2019). In addition, some works (Zhang et al. 2020; Huang et al. 2020) focus on the heatmap-to-coordinate transformation. For example, DARK improves traditional decoding by a Taylor-expansion based refinement (Zhang et al. 2020), while (Huang et al. 2020) introduce a principled unbiased data processing strategy. In a different line of these

*Yan Dai and Xuanhan Wang contribute equally to this work. Lianli Gao is the corresponding author.

[1]https://github.com/vikki-dai/RSGNet

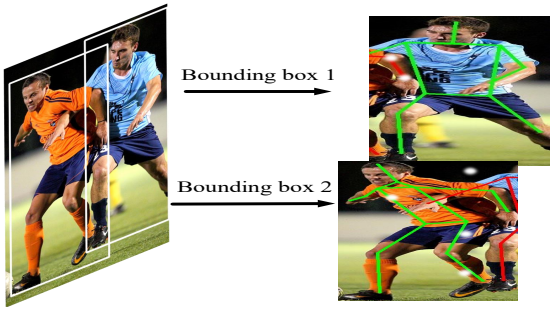

Figure 1: Multi-joints in one bounding box. The green lines denote the pose-of-interest that consists of joints-of-target in a specific bounding box, while the red lines identify pose-of-interference that consists of joints-of-interference. The left hand of person in orange gym shirt is viewed as false positive in the 1st bounding box, while it is set as true positive in the 2nd bounding box.

methods, some works (Moon, Chang, and Lee 2019; Wang et al. 2020a) construct a pipeline of pose 'fixer' for accurate keypoints regression. For example, (Moon, Chang, and Lee 2019) take the estimated pose with errors as input and refine the coordinates in a coarse-to-fine manner. (Wang et al. 2020a) follow this work, and refine the rough localization results through a graph pose refinement mechanism. Compared with bottom-up methods that are detection free, top-down methods often have better pose estimation performance but lower inference speed. Besides, the pipeline of pose 'fixer' consists of three inference steps (e.g., detection, estimation, fixing) which is a huge computation burden from the inference perspective. Instead, we focus on *top-down* methods without any complicated post refinements. Despite the encouraging success that has been achieved in this area, it still remains challenging under the condition of "crowded scenes", where RGB images capture complex real-world scenes with highly-overlapped people, severe occlusions and diverse postures. Specifically, it encounters several problems when directly applying these methods in crowded scenes, as analyzed below:

Multi-joints in one bounding box. The mainstream top-down methods assume that every detected person proposal only contains joints belong to the target person, named joints-of-target. This is impractical especially when under the condition of crowded scenes. As demonstrated in Fig. 1, in addition to joints-of-target, a generated bounding box also contains joints belong to other human instances, named joints-of-interference. Based on above assumption applied in conventional top-down methods, a joint of a person may be assigned different labels. An example of this confusion can also be shown in Fig. 1. When assigning labels to the left hand of person in orange gym shirt, it is set as a positive joint in the second bounding box. However, the left hand becomes a negative one when identifying it in the first bounding box. Moreover, those pose estimators trained on such conditions may mistake joints-of-interference for joints-of-target, and the missing joints-of-target cannot be restored in the post-processing step like pose-NMS. Therefore, joints of a person existing in different proposals should keep label consistency for relieving the confusion of joints estimation model. To achieve this, a straightforward way is to encourage all joints in one bounding box to be active, leading to a multi-joints representation.

Relation Modeling in one bounding box. Once multi-joints from one bounding box have been detected, a joint-to-joint relation modeling method is needed for distinguishing joints-of-target from all detected ones. To address this issue, the recent work (Li et al. 2019) proposes an off-line relation modeling method. Specifically, they first generate all joints within a bounding box, and then a person-joint graph is constructed to distinguish joints-of-target. However, this graph-based approach highly depends on the predicted scores of multi-joints, leading to sub-optimal results. On the other hand, human beings can well identify the joints-of-target according to the human body structure priors. For example, human can easily infer where the 'neck' is after seeing the 'head' and 'shoulder', even though existing other interference 'necks'. Inspired by this, (Qiu et al. 2020) adopt the human body structure priors to enhance the joints features only. However, how to enforce such priors during the process of joints inference is still not explored.

Above challenges motivate us to study two problems: 1) how to design an effective pipeline for crowded scenes pose estimation; and 2) how to equip this pipeline with the ability of relation modeling for interference resolving. To tackle these problems, a multi-joints representation with relation modeling is needed. In this work, we propose a new pipeline named *Relation based Skeleton Graph Network (RSGNet)*. Unlike existing works that directly predict joints-of-target by labeling joints-of-interference as false positive, we first encourage all joints to be predicted and generate a multi-joints heatmap. Next, a Target-aware Relation Parser (TRP) is designed to model the relation over all predicted joints for interference resolving, leading to a target-aware encoding. Furthermore, a Skeleton Graph Machine (SGM) is introduced to model the skeleton-based commonsense knowledge, aiming to enforce the constraint of human body structure during the target pose estimation. Such skeleton-based constraint can help to deal with the challenges in crowded scenes from a reasoning perspective.

To sum up, our work has three main contributions:

- We introduce a new and effective pipeline to tackle the crowded problem of pose estimation, which can be cast as an interference resolution problem. Furthermore, we design a **target-aware relation parser** to model the relation of human joints for interference removal. To the best of our knowledge, this is the first attempt to learn the joints relation in top-down pipeline.
- Inspired by the human beings' inferring ability to identify the joints-of-target according to the human body structure priors, we propose a **skeleton graph machine** to enforce the constraint of human body structure during the process of joints inference for accurate pose estimation.
- Solid experiments show that our proposed ***relation based skeleton graph network (RSGNet)*** significantly outperforms current state-of-the-art pose estimation methods.

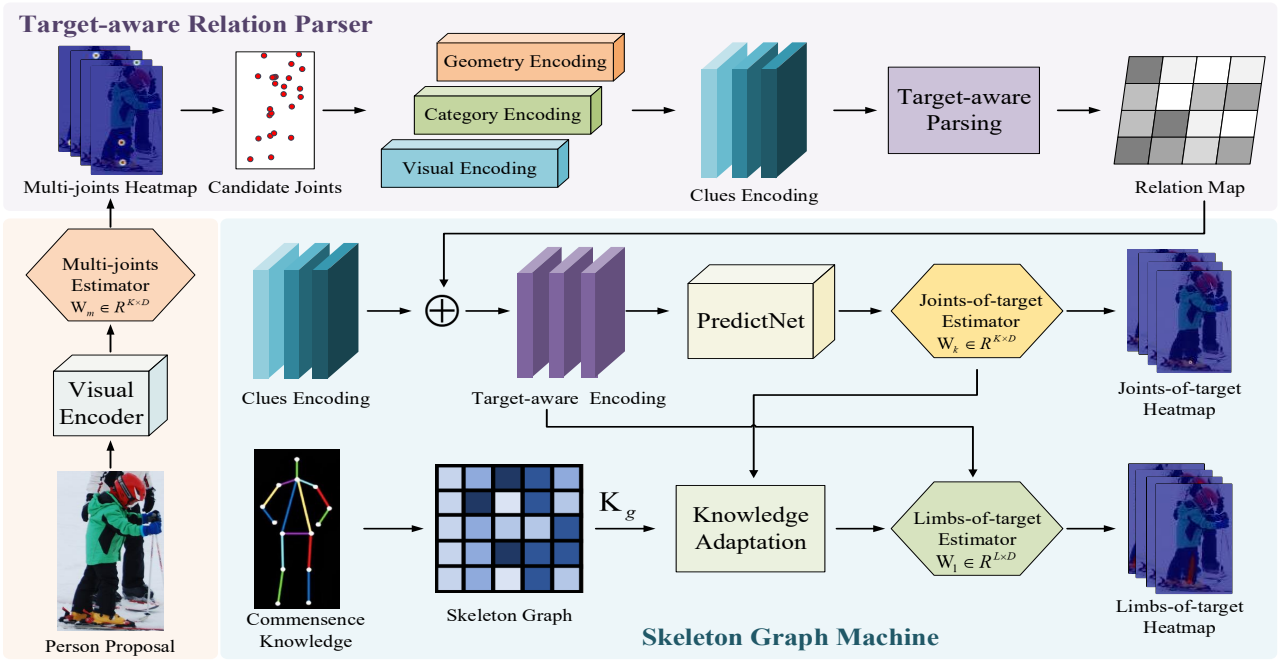

Figure 2: The framework of our RSGNet. The CNN based visual encoder extracts visual features from a region image that contains a target person instance with interference joints, resulting a multi-joints heatmap. The Target-aware Relation Parser then models their relations among all extracted candidate joints by utilizing clues encoding, resulting in a target-aware encoding. Moreover, the Skeleton Graph Machine enforces the skeleton-based commonsense knowledge for target pose estimation.

Proposed Model

Approach Overview

Problem Formulation. Top-down methods handle the pose estimation problem in a two-stage process manner, where it detects all person instances from an image and consecutively performs single person pose estimation for each instance. Given an image of size $H \times W \times 3$, an object detector is adopted to predict N person instances, resulting in N cropped region images $I = \{I_r \in R^{H_r \times W_r \times 3}\}_{r=1}^{N}$. Next, the problem of single person estimation is transformed to estimate K heatmaps $H \in R^{K \times H_h \times W_h}$ of size $H_h \times W_h$, where each heatmap indicates the pixel probability of the corresponding joint category. Finally, these predicted heatmaps will be decoded into a coordinates set $P \in R^{K \times 2}$ of K human joints.

Overview. We follow this popular two-stage pipeline and focus on the second stage. As illustrated in Fig. 2, human region images I predicted by a human detector are fed into the proposed Relation based Skeleton Graph Network (RSGNet) for single person pose estimations. In particular, the proposed method consists of three main components. First, a convolutional neural network (CNN) based visual encoder is applied to extract visual features $f = \{f_r \in R^{H_r \times W_r \times d}\}_{r=1}^{N}$ from region images I. Second, a target-aware relation parser (TRP) is designed to locate candidate joints. And then, it is used to generate a joint-to-joint relation map for interference resolving, resulting in a target-aware encoding. Third, a skeleton graph machine (SGM) is used to infer locations of joints-of-target by utilizing the human body structure priors. In the following sections, we present details of each component.

Target-aware Relation Parser

Given a human region proposal I_r with a rectangle shape, it inevitably contains multiple joints, involving joints-of-target and joints-of-interference, especially in crowded scenes. To address this issue, human beings can distinguish joints-of-target from interferences by utilizing three types of human-centered clues. Besides visual appearance, humans can infer joints-of-target based on another two clues: 1) joint semantics that indicate the difference of inter-classes (e.g., *person* ***vs*** *background*); and 2) joint locations that reflect the unique geometry relation between joints and target person, due to the fact that the bounding box is centered on the target person. Inspired by this, we propose a target-aware relation parser (TRP) for target enhancing and interference resolving. In details, the TRP modular consists of two steps: 1) Clues Encoding; and 2) Target-aware Encoding.

Clues Encoding. Given visual features f_r of the corresponding region image I_r, we first estimate multi-joints heatmap $H_m = \Psi_m(f_r, W_m)$ from them, where Ψ_m is the pixel-wise multi-joints estimator with parameters W_m. Then, a set of N_p candidate joints $P_m = \{P_i\}_{i=1}^{N_p}$ can be generated by thresholding multi-joints heatmap H_m. Different region images results in different number N_p. For each candidate joint, we are given three types of information $\{b_i, c_i, v_i\}_{i=1}^{N_p}$, where $b_i = (\Delta x_i, \Delta y_i, x_i, y_i)$ refers to the joint location.

In particular, the (x_i, y_i) are coordinates and the $(\Delta x_i, \Delta y_i)$ are offsets between the joint and center point of human body. Moreover, c_i is the probability distribution of joint categories estimated by the multi-joints estimator Ψ_m, and v_i denotes the joint visual appearance features extracted from f_r at its location (x_i, y_i). Next, we perform a linear transformation to convert these clues into feature vectors of d dimensions, resulting in three clue encodings: 1) geometry encoding $b^e \in R^{N_p \times d}$; 2) category encoding $c^e \in R^{N_p \times d}$; and 3) visual encoding $v^e \in R^{N_p \times d}$. The final clues encoding $E_c \in R^{N_p \times 3d}$ is constructed by concatenating all three types of clue encodings.

Target-aware Encoding. Given the clues encoding E_c, our goal is to enhance the encodings of joints-of-target and suppress interference ones. To achieve this, we first construct a joint-to-joint relation encoding $E_r \in R^{N_p \times N_p \times d}$ as:

$$E_r \quad = \quad \phi(E_c V^T) \odot \phi(E_c U^T) \tag{1}$$

where $V \in R^{d \times 3d}$ and $U \in R^{d \times 3d}$ are linear transform matrices. ϕ is the ReLU nonlinear activation and $\odot$ denotes the Hadamard product (broadcast element-wise multiplication). Then, a joint-to-joint relation map $A \in R^{N_p \times N_p}$ is generated by a linear function $\Psi_a(E_r, W_r)$ followed by the sigmoid activation, where $W_r \in R^{d \times 1}$ is the linear transform parameters. Each element $A_{i,j}$ represents the probability that candidate joint i and j are grouped to the target person. After that, the relation scores are assigned to candidate joints. Specifically, the target-aware encoding $E_t \in R^{H_h \times W_h \times 3d}$ can be obtained through Equ. 2:

$$E_t \quad = \quad \psi(A E_c) \tag{2}$$

where $\psi(\cdot)$ is the interpolation function that interpolates the enhanced clues encoding back to the heatmap scale $H_h \times W_h$ by the bilinear interpolation. Therefore, the possible target joints (with higher relation score) will be enhanced, while others (lower score) will be suppressed.

Skeleton Graph Machine

As analyzed in *Introduction* section, joints estimation results should satisfy the commonsense knowledge about human body structure priors. In this section, we introduce a skeleton graph machine (SGM) to enforce the constraint of human skeleton priors during the process of joints inference for accurate pose estimation. Specifically, the SGM modular consists of two steps: 1) Skeleton Graph Generation; and 2) Knowledge Adaptation.

Skeleton Graph Generation. Given the target-aware encoding E_t, we can estimate K joints heatmaps $H_t \in R^{K \times H_h \times W_h}$ by applying a joints-of-target estimator Ψ_t with parameters $W_k \in R^{K \times D}$ of size $K \times D$. However, estimating accurate locations of human joints requires the most relative priors knowledge about human body structure. For example, the accurate location of human left arm can help to infer the locations of two related joints (i.e., left elbow and left shoulder). According to such commonsense knowledge, the estimations of human joints should be consistent with the predictions of corresponding human limbs. Therefore, we need to create a skeleton-based graph that provides relation information among joints categories and limbs categories. In particular, we define a joints-to-limbs undirected relation graph $G : G =< C_K, C_L, E >$, where C_K represents human joints with K nodes and C_L denotes human limbs with L nodes. E is the set of edges, each of which connects a node from joints to the one of limbs and encodes a kind of relation knowledge between these two nodes. Such graph can be represented by an adjacent matrix $\mathcal{K}_g \in R^{K \times L}$. For assigning values to each edge, we introduce an identification function $\mathcal{I}(s, k)$, where it outputs 1 *iff* the limb category and the joint category exist a "*has*" relationship (e.g., the left arm limb has the left elbow joint), as formalized in Equ.3:

$$\mathcal{K}_g\,(i, j) = \quad \mathcal{I}(i, j) \tag{3}$$

Given the D dimensional parameters of joints-of-target estimator $W_k \in R^{K \times D}$, we can obtain the relevant parsing parameters $W_s = \mathcal{K}_g^T W_k, \rightarrow R^{L \times D}$ with a matrix multiplication operator.

Knowledge Adaptation. Given the relevant parsing parameters W_s, we need to transform them into parameters of limbs-of-target estimator. To achieve this, we build a small network with two linear functions followed by a LeakyReLU activation. Formally, the parameters of limbs-of-target estimator $W_l \in R^{L \times D}$ are calculated as:

$$W_l \quad = \quad \sigma\left(W_s W_l^1\right) W_l^2 \tag{4}$$

where $W_l^1 \in R^{D \times D}$ and $W_l^2 \in R^{D \times D}$ are linear transform matrices, while $\sigma(\cdot)$ is the LeakyReLU nonlinear function.

The limbs-of-target heatmap $H_l \in R^{L \times H_h \times W_h}$, can be estimated from target encodings E_t through a limbs-of-target estimator with parameters W_l that transferred from joints-of-target parsing parameters W_k. During training, two learning losses (joints and limbs) affect joints parsing parameters, resulting in consistency between joints estimations and limbs estimations. The joints estimation results, therefore, can be constrained by human body structure priors.

Learning Objectives

In this section, we design our learning objectives to enable the proposed model to perform joints estimation in crowded scenes. Given a human-centered region proposal, we input its region into our proposed model and obtain four types of outputs: 1) multi-joints heatmap H_m; 2) joints-of-target heatmap H_t; 3) relation map A; and 4) limbs-of-target heatmap H_l. More specifically, our goal is to enhance joints-of-target responses in joints-of-target heatmap H_t, while encourage all joints to be active in multi-joints heatmap H_m. To achieve this, we follow the previous works (Sun et al. 2019; Li et al. 2019) and use *mean square error (MSE)* as our learning objectives, as defined in Equ. 5:

$$\begin{aligned} \ell_t \quad &= \quad \tfrac{1}{K} \sum_{i=1}^{K} MSE(H_t^i, \hat{H}_t^i) \\ \ell_m \quad &= \quad \tfrac{1}{K} \sum_{i=1}^{K} MSE(H_m^i, \hat{H}_m^i) \end{aligned} \tag{5}$$

where $\hat{H}_t$ and $\hat{H}_m$ are the ground truth of joints-of-target heatmap and multi-joints heatmap, respectively. In particular, $\hat{H}_t$ consists of single-peak Gaussian distribution for

joints-of-target only, while $\hat{H}_m$ consists of multi-peak Gaussian distributions for both joints-of-target and joints-of-interference.

As for joint-to-joint relation learning objective, we take the same strategy and calculate *mean square error (MSE)* between the relation map and its ground truth as below:

$$\ell_r = \frac{1}{N_p^2} \sum MSE(A, \hat{A}) \quad (6)$$

where $\hat{A}$ is the ground truth of the relation map. Notably, $\hat{A}$ contains either 0 or 1. Each element $\hat{A}_{i,j}$ is 1 *iff* i-th joint and j-th joint are grouped into the target person.

As for limbs-of-target learning objective, we calculate *binary cross entropy (BCE)* between the predicted limbs-of-target heatmap H_l and its ground truth through Equ. 7:

$$\ell_l = \frac{1}{L} \sum_{i=1}^{L} BCE(H_l^i, \hat{H}_l^i) \quad (7)$$

where $\hat{H}_l$ is the ground truth of limbs-of-target heatmap and it contains either 0 or 1. Each element $\hat{H}_l(x, y)$ is 1 *iff* the location (x, y) is part of corresponding limb.

Thus, the final learning objective of the whole model can be written as:

$$\mathcal{L} = \alpha\ell_t + \beta\ell_m + \theta\ell_r + \gamma\ell_l \quad (8)$$

where α, β, θ and γ are hyperparameters, and respectively set to 1, 1, 1 and 0.01 for balancing training.

Experiment

In this section, we first introduce datasets and our implementation details. Then, we report our results and comparisons with state-of-the-art approaches. Next, we conduct ablation studies on the proposed components in our method and provide qualitative results with visualization analysis.

Datasets

CrowdPose. We conduct validation experiments on the recently introduced CrowdPose dataset (Li et al. 2019). It contains 20K images and 80k human annotations in total. For each human instance, it is annotated with 14 human joints. Moreover, it is split into two subsets: a training set and a test set with 12K images and 8K images, respectively. Following the previous work (Li et al. 2019), we divide the CrowdPose dataset into three crowding levels by *crowd index* and report standard mean average precision over the test set: 1) mAP (the mean of AP scores at a number of object keypoints similarity (OKS) ranging from 0.5 to 0.95); 2) AP^{Easy} for objects with easy crowding level; 3) AP^{Medium} for objects with medium crowding level; 4) AP^{Hard} for objects with hard crowding level.

MSCOCO. We also report our results on MSCOCO dataset (Lin et al. 2014). It contains over 60K images and 250K person instances annotated with 17 human joints. Moreover, it is divided into three subsets: 1) train2017 with 57K images and 150K person instances; 2) val2017 with 5K images; 3) test-dev2017 with 20K images. We report standard mean average precision on the val set and test-dev set: 1) mAP (the mean of AP scores at a number of object keypoints similarity (OKS) ranging from 0.5 to 0.95); 2) AP^M for objects with medium size; 3) AP^L for objects with large size.

Implementation Details

Training. We adopt Adam optimization algorithm to learn the network parameters, where the batch size is set to 32. The initial learning rate is set to 1e-3 and decays in subsequent iterations. Following the previous work (Sun et al. 2019), we extend the human proposal to a fixed aspect ratio (i.e., height : width = 4 : 3), and crop the region from the image. The cropped region image is further resized to a fixed size (e.g., 256×192). Data augmentation techniques, such as random rotation, random scale and random flipping, are used during training for reducing the risk of over-fitting.

Inference. Our method follows the top-down framework: detects person first, and then performs single person estimation. Since the human detector is not what we focus on, we simply adopt ResNet101-FPN (Lin et al. 2017) as our human detector and re-train it on CrowdPose dataset. As for COCO dataset, we simply use the same person detector provided by HRNet (Sun et al. 2019) for fair comparison. For heatmap-to-coordinates decoding method, we follow the previous works (Sun et al. 2019; Xiao, Wu, and Wei 2018; Li et al. 2019), and simply compute the final joints-of-target heatmap by averaging the heatmaps of the original and flipped images. The coordinates of joints are adjusted from the heatmap's highest value location.

We implement our method based on PyTorch deep learning library on a server with 4 NVIDIA RTX GPUs, and adopt HRNet (Sun et al. 2019) as our backbone for all experiments. Other details are identical as details in HRNet (Sun et al. 2019). More network design details are seen in *supplementary materials*.

Comparison with State-of-the-art Methods

Comparisons on CrowdPose Dataset. Quantitative results of the proposed method and current state-of-the-art methods on CrowdPose *test* set are listed in Tab. 1. As shown in Tab. 1, our RSGNet with HRNet-W32 achieves the best performance in all the measure metrics, outperforming other methods with the same input size or same backbone. Compared to HRNet (Sun et al. 2019), the proposed RSGNet, with the same input size and the same backbone, achieves 1.9%, 0.8%, 1.3% and 0.5% gains, respectively. When compared to previous best-performed OPEC-Net (Qiu et al. 2020), our method with the small backbone of HRNet-w32 and the input size of 256×192, can obtain significant improvements, reaching 3.0% gains.

Comparisons on MSCOCO Dataset. We also evaluate our method with current state-of-the-art methods on COCO *val* set. The results are given in Tab. 2. Without bells and whistles, our RSGNet with HRNet-W32 backbone achieves 75.7% mAP. More specifically, our approach improves HRNet-32 by 1.3% mAP (74.4 to 75.7) and 0.9% mAP (75.1 to 76.0) when input region images with the size of 256×192 and 384×288, respectively. Similar gains (0.8% and 0.6%) can also be observed when applying HRNet-W48 backbone. This shows that our method can perform improvement on general pose estimation problem as well.

We also evaluate our method on the *test-dev* set and the results are summarized in Tab. 3. As Tab. 3 reports, our method

Method	Backbone	Input size	AP	AP^{50}	AP^{75}	AP^{Easy}	AP^{Medium}	AP^{Hard}
Bottom-up methods								
OpenPose(Cao et al. 2018)	CPM	-	-	-	-	62.7	48.7	32.3
HihgerHRNet (Cheng et al. 2020)	HRNet-W48	-	67.6	87.4	72.6	75.8	68.1	58.9
Top-down methods								
Mask-RCNN (He et al. 2017)	ResNet-101	-	57.2	83.5	60.3	69.4	57.9	45.8
SimpleBaseline (Xiao, Wu, and Wei 2018)	ResNet-50	256×192	60.8	81.4	65.7	67.3	86.3	71.8
AlphaPose (Li et al. 2019)	ResNet-101	320×256	66.0	84.2	71.5	75.5	66.3	57.4
OPEC-Net (Qiu et al. 2020)	ResNet-101	320×256	70.6	86.8	75.6	-	-	-
HRNet (Sun et al. 2019)	HRNet-W32	256×192	71.7	89.8	76.9	79.6	72.7	61.5
RSGNet (**Ours**)	HRNet-W32	256×192	**73.6 (+1.9)**	**90.7**	**79.0**	**81.3**	**74.6**	**63.4**
HRNet (Sun et al. 2019)	HRNet-W32	384×288	73.5	90.7	78.9	81.2	74.5	63.2
RSGNet (**Ours**)	HRNet-W32	384×288	**74.3 (+0.8)**	**90.7**	**79.7**	**81.8**	**75.3**	**64.6**
HRNet (Sun et al. 2019)	HRNet-W48	256×192	73.3	90.0	78.7	81.0	74.4	63.4
RSGNet (**Ours**)	HRNet-W48	256×192	**74.6 (+1.3)**	**90.9**	**80.1**	**82.0**	**75.6**	**64.5**

Table 1: Comparison with the state-of-the-art methods on CrowdPose *test* dataset.

Method	Backbone	Input size	# Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
CPN (Chen et al. 2018)	ResNet-50	256×192	27.0M	6.20	68.6	-	-	-	-	-
SimpleBaseline (Xiao, Wu, and Wei 2018)	ResNet-152	256×192	68.6M	15.7	72.0	89.3	79.8	68.7	78.9	77.8
HRNet (Sun et al. 2019)	HRNet-W32	256×192	28.5M	7.10	74.4	90.5	81.9	70.8	81.0	79.8
RSGNet (**Ours**)	HRNet-W32	256×192	29.2M	8.31	**75.7 (+1.3)**	**90.5**	**82.0**	**71.8**	**82.5**	**80.8**
HRNet (Sun et al. 2019)	HRNet-W32	384×288	28.5M	16.0	75.8	90.6	82.7	71.9	82.8	81.0
RSGNet (**Ours**)	HRNet-W32	384×288	29.2M	18.7	**76.6 (+0.8)**	**91.0**	**82.9**	**72.8**	**83.3**	**81.6**
HRNet (Sun et al. 2019)	HRNet-W48	256×192	63.6M	14.6	75.1	90.6	82.2	71.5	81.8	80.4
RSGNet (**Ours**)	HRNet-W48	256×192	64.5M	16.9	**76.0 (+0.9)**	**90.8**	**82.6**	**72.1**	**82.9**	**81.1**
HRNet (Sun et al. 2019)	HRNet-W48	384×288	63.6M	32.9	76.3	90.8	82.9	72.3	83.4	81.2
RSGNet (**Ours**)	HRNet-W48	384×288	64.5M	38.0	**77.0 (+0.7)**	**91.0**	**83.6**	**72.9**	**83.9**	**81.7**

Table 2: Comparison with the state-of-the-art methods on COCO *val* dataset.

Method	Backbone	Input size	# Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Mask-RCNN (He et al. 2017)	ResNet-50	-	-	-	63.1	87.3	68.7	57.8	71.4	-
CPN (Chen et al. 2018)	ResNet-152	384×288	-	-	72.1	91.4	80.0	68.7	77.2	78.5
AlphaPose (Fang et al. 2017)	PyraNet	320×256	28.1M	26.7	72.3	89.2	79.1	68.0	78.6	-
Posefix (Moon, Chang, and Lee 2019)	ResNet-152	384×288	68.6M	35.6	73.6	90.8	81.0	70.3	79.8	79.0
OPEC-Net (Qiu et al. 2020)	ResNet-101	320×256	-	-	73.9	91.9	82.2	-	-	-
SimpleBaseline (Xiao, Wu, and Wei 2018)	ResNet-152	384×288	68.6M	35.6	73.7	91.9	81.1	70.3	80.0	79.0
HRNet (Sun et al. 2019)	HRNet-W32	256×192	28.5M	7.10	73.5	92.2	81.9	70.2	79.2	79.0
RSGNet (**Ours**)	HRNet-W32	256×192	29.2M	8.31	**74.7 (+1.2)**	**92.3**	**82.3**	**71.4**	**80.5**	**79.9**
HRNet (Sun et al. 2019)	HRNet-W32	384×288	28.5M	16.0	74.9	92.5	82.8	71.3	80.9	80.1
RSGNet (**Ours**)	HRNet-W32	384×288	29.2M	18.7	**75.7 (+0.8)**	**92.5**	**83.1**	**71.9**	**81.7**	**80.9**
HRNet (Sun et al. 2019)	HRNet-W48	256×192	63.6M	14.6	74.3	**92.4**	82.6	71.2	79.6	79.7
RSGNet (**Ours**)	HRNet-W48	256×192	64.5M	16.9	**75.1 (+0.8)**	92.3	**82.7**	**71.6**	**80.9**	**80.3**
HRNet (Sun et al. 2019)	HRNet-W48	384×288	63.6M	32.9	75.5	92.5	83.3	71.9	81.5	80.5
RSGNet (**Ours**)	HRNet-W48	384×288	64.5M	38.0	**76.0 (+0.5)**	**92.6**	**83.4**	**72.3**	**82.0**	**81.2**

Table 3: Comparison with the state-of-the-art methods on COCO *test-dev* dataset.

with HRNet-w48 backbone at the input size of 384×288 achieves the best, and it outperforms HRNet-w48 with the same input size by 0.5%. Similar gains also can be observed in other experimental settings, demonstrating the effectiveness and generalizability of our approach.

Ablation Study

In this section, we conduct diagnostic experiments to perform component analysis of our proposed Relation based Skeleton Graph Network (RSGNet). In particular, we use the HRNet-W32 with the input resolution of 256×192 (Sun et al. 2019) as our baseline and compare three different models: 1) HRNet-W32, which is the baseline approach pretrained on ImageNet. 2) HRNet-W32 with the Target-aware Relation Parser (TRP) but without the Skeleton Graph Machine (SGM). In this setting, the final estimation of joints-of-target heatmap is estimated from the target-aware encoding by directly applying the joints-of-target estimator. 3) HRNet-W32 with the TRP module and the SGM module simultaneously, which is our proposed RSGNet. In this setting, we enforce the constraint of human body structure during the target pose estimation. Furthermore, more detailed component analyses for the TRP and SGM modulars are presented in *supplementary materials*.

Component Ablation Studies on CrowdPose. Our ablation study on CrowdPose *test* set from the baseline gradually to all components incorporated is summarized in Tab. 4. From the results, we can observe that the whole framework has the

Figure 3: Qualitative results comparison on CrowdPose *test* set, which across various overlaps with different difficulty levels. The first two rows show some miss-detected joints-of-target while the second two rows show some cases of mistaking joints-of-interference for joints-of-target, when it encounters more complex crowded scenes. The red circles spot the cases that HRNet fails to estimate, while the green circles spot the corresponding estimations produced by our RSGNet.

CrowdPose *test* dataset						
HRNet-w32	TRP	SGM	AP	AP^{Easy}	AP^{Medium}	AP^{Hard}
✓			71.7	79.6	72.7	61.5
✓	✓		73.1	80.9	74.2	62.8
✓	✓	✓	**73.6**	**81.3**	**74.6**	**63.4**
Gains			**+1.9**	**+1.7**	**+1.9**	**+1.9**
COCO *minival* dataset						
HRNet-w32	TRP	SGM	AP	AP^{M}	AP^{L}	AR
✓			74.4	70.8	81.0	79.8
✓	✓		74.9	71.3	81.5	80.1
✓	✓	✓	**75.7**	**71.8**	**82.5**	**80.8**
Gains			**+1.3**	**+1.0**	**+1.5**	**+1.0**

Table 4: Ablation Study. Investigating the effect of proposed modules.

best performance with 73.6% mAP scores, improving baseline model by 1.9%. More specifically, HRNet-W32 with the TRP module outperforms the baseline by 1.4% on mAP metric, 1.3% on easy crowding setting, 1.5% and 1.3% on medium and hard crowding setting, respectively. It indicates that the proposed TRP module can effectively alleviate the effect of interference joints. Furthermore, slight gains can be obtained after enforcing the constraint of human body structure during the target pose estimation, implying the necessity of applying joints-to-limbs constraint.

Component Ablation Studies on COCO. We also gradually add proposed components for ablation studies on COCO *val* set. The results are also shown in Tab. 4. The proposed RSGNet is also improving for COCO dataset, which yields 1.3 improvements in terms of mAP. For more details, it improves AP^M by 1.0%, AP^L by 1.5% and AR by 1.0%. This clearly indicates the effectiveness and generalizability of the proposed RSGNet.

Visualization. We provide qualitative results across various crowded scenes and compare the proposed RSGNet with the HRNet in Fig. 3. As shown in Fig. 3, the first two rows are qualitative comparisons across various overlaps with different difficulty levels. As it can be seen, the HRNet fails to estimate some human joints when the crowded scenes become complex. The second two rows present qualitative comparison for interference resolving. For example, the column 1 shows that the HRNet wrongly connects a woman's right leg to her left ankle due to the complex overlapping, while the RSGNet can well resolve these interferences. Similar cases also can be observed even under higher complexity of crowded scenes (Seen in column 2-5).

Conclusion

In this paper, we propose a novel relation based skeleton graph network (RSGNet) for multi-person pose estimation in crowded scenes. In particular, we design a target-aware relation parser to address the issue of interference resolving. Furthermore, we introduce a skeleton graph machine to enforce the constraint of human body structure during the joints inference for accurate pose estimation. Extensive experiments on CrowdPose and MSCOCO benchmarks prove the effectiveness of our approach.

Acknowledgments

This work is supported National Key Research and Development Program of China (No. 2018AAA0102200), the National Natural Science Foundation of China (Grant No. 61772116, No. 61872064, No. 62020106008), Sichuan Science and Technology Program (Grant No.2019JDTD0005), The Open Project of Zhejiang Lab (Grant No.2019KD0AB05) and Open Project of Key Laboratory of Artificial Intelligence, Ministry of Education (Grant No. AI2019005).

References

Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.; and Sheikh, Y. 2018. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *CoRR* abs/1812.08008.

Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; and Sun, J. 2018. Cascaded Pyramid Network for Multi-Person Pose Estimation. In *CVPR*, 7103–7112.

Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T. S.; and Zhang, L. 2020. HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation. In *CVPR*.

Fang, H.; Lu, G.; Fang, X.; Xie, J.; Tai, Y.; and Lu, C. 2018. Weakly and Semi Supervised Human Body Part Parsing via Pose-Guided Knowledge Transfer. In *CVPR*, 70–78.

Fang, H.; Xie, S.; Tai, Y.; and Lu, C. 2017. RMPE: Regional Multi-person Pose Estimation. In *ICCV*, 2353–2362.

Gao, L.; Li, X.; Song, J.; and Shen, H. T. 2020. Hierarchical LSTMs with Adaptive Attention for Visual Captioning. *IEEE Trans. Pattern Anal. Mach. Intell.* 1112–1131.

Golda, T.; Kalb, T.; Schumann, A.; and Beyerer, J. 2019. Human Pose Estimation for Real-World Crowded Scenarios. In *AVSS*, 1–8.

Guo, Y.; Gao, L.; Song, J.; Wang, P.; Xie, W.; and Shen, H. T. 2019. Adaptive Multi-Path Aggregation for Human DensePose Estimation in the Wild. In *ACM MM*, 356–364.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. B. 2017. Mask R-CNN. In *ICCV*, 2980–2988.

Huang, J.; Zhu, Z.; Guo, F.; and Huang, G. 2020. The Devil Is in the Details: Delving Into Unbiased Data Processing for Human Pose Estimation. In *CVPR*, 5699–5708.

Jin, S.; Liu, W.; Xie, E.; Wang, W.; Qian, C.; Ouyang, W.; and Luo, P. 2020. Differentiable Hierarchical Graph Grouping for Multi-Person Pose Estimation. *CoRR* abs/2007.11864.

Li, J.; Wang, C.; Zhu, H.; Mao, Y.; Fang, H.; and Lu, C. 2019. CrowdPose: Efficient Crowded Scenes Pose Estimation and a New Benchmark. In *CVPR*, 10863–10872.

Lin, T.; Dollár, P.; Girshick, R. B.; He, K.; Hariharan, B.; and Belongie, S. J. 2017. Feature Pyramid Networks for Object Detection. In *CVPR*, 936–944.

Lin, T.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollar, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*, 740–755.

Moon, G.; Chang, J. Y.; and Lee, K. M. 2019. PoseFix: Model-Agnostic General Human Pose Refinement Network. In *CVPR*, 7773–7781.

Newell, A.; Huang, Z.; and Deng, J. 2017. Associative Embedding: End-to-End Learning for Joint Detection and Grouping. In *NIPS*, 2277–2287.

Nie, X.; Feng, J.; Zhang, J.; and Yan, S. 2019. Single-Stage Multi-Person Pose Machines. In *ICCV*, 6950–6959.

Papandreou, G.; Zhu, T.; Chen, L.; Gidaris, S.; Tompson, J.; and Murphy, K. 2018. PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model. *arXiv preprint arXiv:1803.08225* .

Qiu, L.; Zhang, X.; Li, Y.; Li, G.; Wu, X.; Xiong, Z.; Han, X.; and Cui, S. 2020. Peeking into occluded joints: A novel framework for crowd pose estimation. *CoRR* abs/2003.10506.

Shotton, J.; Sharp, T.; Kipman, A.; Fitzgibbon, A. W.; Finocchio, M.; Blake, A.; Cook, M.; and Moore, R. 2013. Real-time human pose recognition in parts from single depth images. *Commun. ACM* 116–124.

Song, J.; Zhang, H.; Li, X.; Gao, L.; Wang, M.; and Hong, R. 2018. Self-Supervised Video Hashing With Hierarchical Binary Auto-Encoder. *IEEE Trans. Image Process.* 3210–3221.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. In *CVPR*, 5693–5703.

Wang, J.; Long, X.; Gao, Y.; Ding, E.; and Wen, S. 2020a. Graph-PCNN: Two Stage Human Pose Estimation with Graph Pose Refinement. *CoRR* abs/2007.10599.

Wang, X.; Gao, L.; Song, J.; and Shen, H. T. 2020b. KTN: Knowledge Transfer Network for Multi-person DensePose Estimation. In *ACM MM*, 3780–3788.

Wang, X.; Gao, L.; Wang, P.; Sun, X.; and Liu, X. 2018. Two-Stream 3-D convNet Fusion for Action Recognition in Videos With Arbitrary Size and Length. *IEEE Trans. Multim.* 634–644.

Xiao, B.; Wu, H.; and Wei, Y. 2018. Simple Baselines for Human Pose Estimation and Tracking. In *ECCV*, 472–487.

Zhang, F.; Zhu, X.; Dai, H.; Ye, M.; and Zhu, C. 2020. Distribution-Aware Coordinate Representation for Human Pose Estimation. In *CVPR*, 7091–7100.