

I^2R -Net: Intra- and Inter-Human Relation Network for Multi-Person Pose Estimation

Yiwei Ding^{1*}, Wenjin Deng^{1*}, Yinglin Zheng¹, Pengfei Liu¹,
Meihong Wang¹, Xuan Cheng¹, Jianmin Bao², Dong Chen², Ming Zeng^{1†}

¹School of Informatics, Xiamen University

²Microsoft Research Asia

{dingyiwei,dengwenjin,zhengyinglin,pengfei}@stu.xmu.edu.cn,

{wangmh,chengxuan,zengming}@xmu.edu.cn,{jianbao, doch}@microsoft.com

Abstract

In this paper, we present the Intra- and Inter-Human Relation Networks (I^2R -Net) for Multi-Person Pose Estimation. It involves two basic modules. First, the Intra-Human Relation Module operates on a single person and aims to capture Intra-Human dependencies. Second, the Inter-Human Relation Module considers the relation between multiple instances and focuses on capturing Inter-Human interactions. The Inter-Human Relation Module can be designed very lightweight by reducing the resolution of feature map, yet learn useful relation information to significantly boost the performance of the Intra-Human Relation Module. Even *without bells and whistles*, our method can compete or outperform current competition winners. We conduct extensive experiments on COCO, CrowdPose, and OCHuman datasets. The results demonstrate that the proposed model surpasses all the state-of-the-art methods. Concretely, the proposed method achieves **77.4%** AP on CrowdPose dataset and **67.8%** AP on OCHuman dataset respectively, outperforming existing methods by a large margin. Additionally, the ablation study and visualization analysis also prove the effectiveness of our model.

1 Introduction

2D Multi-person Pose Estimation (MPPE) aims to detect and localize the human keypoints for all persons appearing in a given image. Since human pose provides abundant structural and motion information, MPPE has attracted attentions in wide applications, such as human activity understanding, human-object interaction, avatar animation, etc.

Current MPPE methods can be divided into two fashions, *i.e.*, top-down and bottom-up. Top-down methods [He *et al.*, 2017; Xiao *et al.*, 2018; Fang *et al.*, 2017; Sun *et al.*, 2019; Chen *et al.*, 2018] detect bounding box for each human, then estimate the pose of each human separately. Bottom-up methods [Cao *et al.*, 2019; Newell *et al.*, 2016a; Cheng *et al.*,

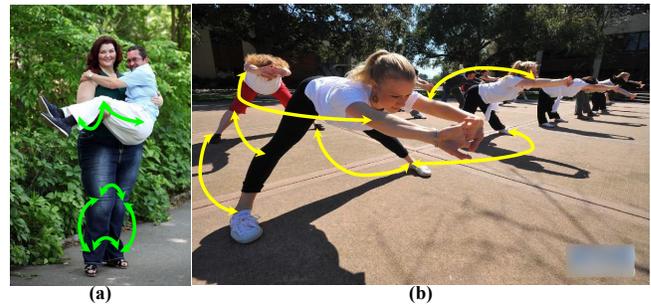


Figure 1: **Illustrations of Intra- and Inter-Human Relations.** (a) the intra-human relation is marked as the green arrow (left-right pairs and two keypoints of the same limb), while the inter-human relation is denoted as the yellow arrow (the same type of keypoints in different individuals).

2020; Papandreou *et al.*, 2018; Jin *et al.*, 2020] localize all joints in the image, and then group them into individuals. Despite having achieved great performance in MPPE, the above methods produce inferior results in crowded scenes, due to overlapping, self-occlusion, and a large variety of appearances and poses. Top-down methods treat each person separately, neglecting the informative cues of mutual interaction among persons. On the other hand, bottom-up methods jointly detect joints of all persons in an image, potentially correlating joints across persons, but the relationships between persons are vague and agnostic. Recently, several pioneers address the above challenges via decoupling poses [Li *et al.*, 2019; Qiu *et al.*, 2020] or fusing multiple possible poses [Wang *et al.*, 2021a] at a single person. However, these methods still do not fully take the multiple-human correlations into account, leaving these important interactions unexplored.

This work investigates how to leverage the correlations of intra-human or cross-human, *e.g.*, persons with similar poses or persons who closely interact with each other, to improve the accuracy of multi-person pose estimation. As shown in Figure 1(a), two people are hugging each other. Their poses show strong intra-human information (*e.g.*, left-right pairs and two keypoints of the same limb) and obvious inter-human relations. In Figure 1(b), these persons are dancing together, indicating that their poses are similar, also providing inter-human cues for more accurate pose estimation, especially in

*Equal Contribution.

†Corresponding Author.

the occluded regions or low-resolution regions.

Based on above observations, we propose a novel two-stage Intra- and Inter-Human Relation Networks (I^2R -Net) for MPPE. The first stage targets at learning the dependencies among body parts of a single person. It operates at high resolution and generates the locations of small parts, such as the eyes, wrists, and ankles. The second stage operates multi-person collaboratively at low resolution. It aims to capture interactions between instances, like touching hands, joining arms, and crossing legs. It's worth noting that each instance is represented by a low-resolution feature map rather than merely a vector. By doing this, we can obtain variant responses from regions, since the cross-instance interaction depends on spatial semantic information. Meanwhile, the low-resolution feature map significantly reduces the computation cost of the second stage but is sufficient to capture useful correlation information. We call the first **Intra-Human Relation Module**, and the second **Inter-Human Relation Module**. The two stages are sequentially stacked with skip-connection.

This two-stage framework is structure flexible and function specialized for MPPE. The Intra-Human Relation Module can be an arbitrary single-person pose estimation method, aiming to explore part-level patterns for keypoint detection. Moreover, the Inter-Human Relation Module can be implemented by any non-local model, to pay more attention to correlations in semantic features. The two modules are able to work collaboratively to infer multi-person poses, because the first module offers high-quality pose information to the next module, while the second module aids the first to eliminate ambiguities in occluded or less discriminative regions.

We evaluate our method on common benchmarks for crowded scenes including CrowdPose and OCHuman. Experiments have proved that our method surpasses all state-of-the-art approaches by large margins. We achieve 77.4% AP on CrowdPose, outperforming HRFormer-B by 5.0%. We achieve 67.8% AP on OCHuman dataset, outperforming TransPose-H by 5.5%. In addition, we bring improvements on COCO dataset, which means our method can generally work in non-crowded scenes.

In summary, the contributions of our work are as follows:

- We propose a two-stage MPPE framework that not only relates each body part in a single person, but also builds connections among multi-person in the images. The proposed method bridges the gap between pose estimation and human-human interactions.
- The framework is designed to be flexible. The module coping with intra-human relations can be arbitrary pose estimation method. While the module that models inter-human relations is very lightweight.
- Extensive experiments show that, without bells and whistles, our method significantly surpasses state-of-the-art methods on challenging datasets like CrowdPose, OCHuman, and COCO.

2 Related Work

2.1 Multi-Person Pose Estimation

Top-down Framework. Thanks to the great success of object detection, it is intuitive to estimate the pose of each person with the given human bounding box. Thus, lots of work under top-down framework focus on developing single-person estimation without optimizing the off-the-shelf detection models. Reviewing the methods in recent years, the major development idea is to improve the feature representation in spatial. [He *et al.*, 2017] directly adds a keypoint detection branch on a CNN-based feature extractor. [Newell *et al.*, 2016b] follows multi-stage fashion to stack modules for dense prediction. Furthermore, applied with intermediate supervision at each stage, the model is driven to perform estimation from coarse to fine. CPN [Chen *et al.*, 2018] employs pyramid structure in spatial to handle parts on different scales. Instead of deepening network with sequential modules, HRNet [Sun *et al.*, 2019] maintains multi-resolution in parallel leading to rich high-resolution representation.

Bottom-up Framework. Bottom-up methods first predict all identity-free body keypoints from an uncropped image, then group these body keypoints into different individuals. Most of them adopt heatmap for keypoint detection, and focus on how to identify and group keypoints to the corresponding person. OpenPose [Cao *et al.*, 2019] proposed the part affinity field to reveal the relationship between two keypoints in the same limb. During the grouping stage, after calculating the line integral score between two candidate keypoints, the pair with the highest score is associated. This approach is denoted as PAF family. Another popular fashion is Associative Embedding, called AE [Newell *et al.*, 2016a; Cheng *et al.*, 2020]. They learn the tag embeddings for each person and group keypoints into individuals by clustering the tag.

Top-down methods estimate poses one by one, without taking the relationship between instances into account. While bottom-up methods detect all poses together, the relationship cues are vague. In contrast, our work pays attention to capturing the hierarchical interactions in the multi-person scenario to promote pose estimation performance. Thus, we provide a novel two-stage framework to focus on modeling part-level and instance-level relationships for MPPE.

Human Pose Estimation in Crowd. The performance of previous work decrease as the number of human increases. Even though top-down methods achieve higher performance than bottom-up methods, they still fail in crowded scenes due to occlusions. Thus, CrowdPose dataset [Li *et al.*, 2019] and OCHuman dataset [Zhang *et al.*, 2019] have been proposed to encourage researchers to study this challenging problem. Existing methods [Li *et al.*, 2019; Qiu *et al.*, 2020] attempt to decompose coupled poses in crowd.

In this paper, however, the proposed Inter-Human Relation Module is designed to conquer this dilemma by leveraging the relationships between instances.

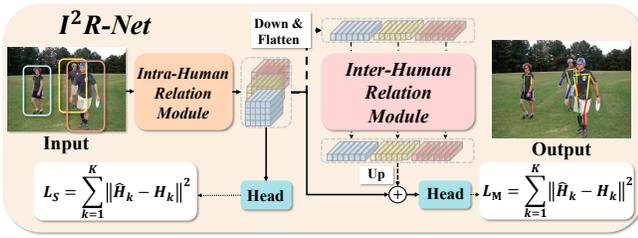


Figure 2: I^2R -Net Framework. The framework includes an Intra-Human Relation Module and an Inter-Human Relation Module. The Intra-Human Relation Module can apply current single-person pose estimation methods flexibly. The Inter-Human Relation Module is lightweight by downsampling the feature from Intra-Human Relation Modules.

2.2 Vision Transformer

Transformer is an attention-based model, and has been studied in Neural Language Processing in recent years, like BERT [Devlin *et al.*, 2018] and GPT [Radford *et al.*, 2019]. With the effectiveness of Transformer, recently, an increasing number of works [Parmar *et al.*, 2018; Bello *et al.*, 2019; Touvron *et al.*, 2021] use Vision Transformer for computer vision tasks. Several previous works apply Transformer to 2D pose estimation. For example, Transpose [Yang *et al.*, 2021] proposes a pose estimator with attention layers to capture and reveal the spatial dependencies between 2D heatmaps. HRFormer [Yuan *et al.*, 2021] learns high-resolution representations via self-attention schema. However, none of them take human relations into account.

Although recent work [Wang *et al.*, 2021b; Mihai *et al.*, 2021] utilize Transformer to calculate relations, yet they aim to solve 3d motion prediction or reconstruction where the relationships between different persons are intuitive. In contrast, our work is the first to incorporate multi-person correlations in single image 2D Pose Estimation, where the relationship is hard to build and we propose a novel module to solve this. Moreover, instead of using a well-designed framework for only building inter-human relationships, we propose a flexible, lightweight, and general module for this purpose that can be easily integrated into current or future single-person pose estimation frameworks.

3 Approach

3.1 Overview of I^2R -Net

As shown in Figure 2, the framework takes the detected and cropped human image patches in the image as input. In the first stage, Intra-Human Relation Network (Section 3.2) extracts the feature of each human patch separately and preserves relatively high-resolution feature representation. In the second stage, Inter-Human Relation Module (Section 3.3) exchanges information of the per-person features extracted by the first stage, and the per-person features should be downsampled beforehand to reduce computation cost. Finally, the two stages are stacked with a skip-connection to fuse the high-resolution intra-human information and the low-resolution inter-human information.

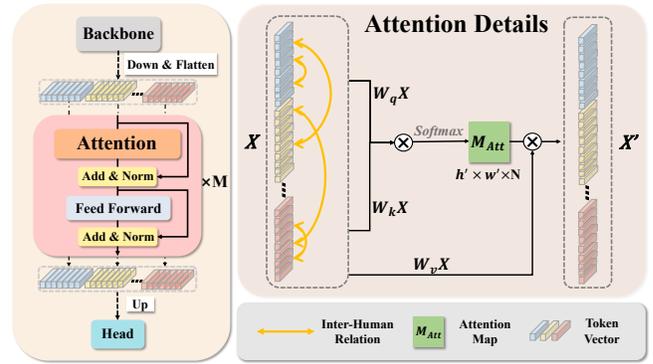


Figure 3: A Vanilla Inter-Human Network. The vanilla network uses Inter-Human Relation Module only with a conventional CNN-based model to extract features of each person cropped from the input image. The extracted features are downsampled, flattened, and concatenated before being passed into Inter-Human Relation Module. The output features of the Inter-Human Relation Module are upsampled to higher resolution, and then fed into a keypoint prediction head.

3.2 Intra-Human Relation Module

As the goal of Intra-Human Relation Module is to learn semantic representations of the human patches and capture the dependencies between body parts, it operates at patches of cropped human images. It can be an arbitrary single-person human pose estimation method. One key factor is the input and output resolutions, which is a trade-off between performance and computational consumption. Following the recent works [Yang *et al.*, 2021; Yuan *et al.*, 2021], we adopt two settings, *i.e.*, 256×192 of input and 64×48 of output as default, and 384×288 of input and 96×72 of output for comparison.

3.3 Inter-Human Relation Module

For the task of Multi-Human Pose Estimation, we devise an Inter-Human Relation Module to collaboratively infer the pose of each person via explicitly modeling cross-human correlation.

The Inter-Human Relation Module is lightweight yet effective. We illustrate a vanilla network example in Figure 3. Given an image I with N persons, the Intra-Human Relation Module (*e.g.*, HRNet, TransPose) locates all persons and extracts their features $\mathbf{P}_1, \mathbf{P}_2, \dots, \mathbf{P}_N$ from raw patches. These features contain pose information of each person. The Inter-Human Relation Module takes these features as input and exchanges pose information among each instance. To fully capture the correlations between different people, we adopt the recently popular Transformer blocks. It is worth noting that, as shown in Figure 3, the Inter-Human relation module not only exchanges information across instances, but also correlates information among parts in each instance. This ensures that even using only an Inter-Human Relation Module (without the Intra- stage, and the input per-person features are extracted by a backbone), it can still achieve state-of-the-art performance in our experiments.

Concretely, suppose the output feature maps of first stage is $\mathbf{P}_i \in \mathbb{R}^{h \times w \times d}$, $i \in [1, N]$, where $h \times w$ is the feature map

resolution, d is the channel dimension, N is the number of persons. We use the max pooling operator with kernel size and stride set to R to operate on \mathbf{P}_i and get resulted down-sampled feature map $\mathbf{P}_i' \in \mathbb{R}^{h/R \times w/R \times d}$. Then we flatten all the features in \mathbf{P}_i' and concatenate them to get a sequence with a length of $L = h/R \times w/R \times N$:

$$\mathbf{P}_i' = \text{Downsample}(\mathbf{P}_i) \quad (1)$$

$$\mathbf{X} = \text{Concat}[\text{Flatten}(\mathbf{P}_i')], i \in [1, N] \quad (2)$$

where $\mathbf{X} \in \mathbb{R}^{L \times d}$.

Subsequently, \mathbf{X} is fed into M (set to 6 for vanilla network) stacked Transformer blocks, aiming to compute correlations between different instances. The formulation of attention operation is given as follows,

$$\text{Attention}(X) = \text{Softmax} \left[\frac{(\mathbf{W}_q \mathbf{X})(\mathbf{W}_k \mathbf{X})^T}{\sqrt{d}} \right] \mathbf{W}_v \mathbf{X} \quad (3)$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d \times d}$ are learnable parameters.

Finally, we use a standard deconvolution layer to upsample the output embeddings \mathbf{X}' to higher resolution and employ a keypoint head to predict K types of keypoint heatmaps $\hat{\mathbf{H}}$, *i.e.*, $\hat{\mathbf{H}} = \text{Head}(\text{Upsample}(\mathbf{X}'))$.

Following the previous methods, we adopt the MSE loss to supervise predicted keypoints heatmaps $\hat{\mathbf{H}}$ with the ground-truth \mathbf{H} , *i.e.*, $\sum_{k=1}^K \|\hat{\mathbf{H}}_k - \mathbf{H}_k\|^2$. In our framework, the Intra-Human loss \mathcal{L}_S serves as an intermediate supervision. Thus, after summing up instance-level stage loss \mathcal{L}_M , the total loss is $\mathcal{L} = \mathcal{L}_S + \alpha \mathcal{L}_M$. In our experiments, we set $\alpha = 1$.

Our proposed Inter-Human Relation Module is general and flexible, and it can be instantiated with different attention methods (*e.g.*, Non-local networks, Transformer Layers) and implementation specifics.

Input of Inter-Human Relation Module. To reduce the computation complexity, we downsample the output feature map of the Intra-Human Relation module with the ratio R and restrict the number of persons to N . Suppose the input feature map size is $h \times w$, the downsampled feature map is $h' \times w'$, then we flatten and concatenate all these elements together. After that, we perform self-attention calculations on all these feature elements (shown as Figure 3). The correlations between poses of different persons are captured. Since N is an important hyper-parameter for the computation complexity. We set a max person number N for each input image differently on different datasets. Specifically, we adopt $N = 6$ for CrowdPose, $N = 3$ for OCHuman and $N = 4$ for COCO, respectively. Experimental results prove that our input setting of the Inter-Human Relation Module achieves a strong performance while introducing little extra computation cost.

Person Selection Mechanism. In fact, the number of people usually varies in images. To this end, we introduce the person selection mechanism. As mentioned above, we set N as the maximum number of person patches.

During the training phase, we use the patch selection mechanism to facilitate the network training. For cases with persons less than N in an input image, we simply calculate correlations between these persons or even a single person. But for parallel computing, we pad noise features to increase the person number to N and mask them out to prevent self-attention calculation in Transformer blocks. When the number exceeds N , we randomly select a person as the target and fill the patches with the $N - 1$ nearest neighbors.

In the testing stage, we apply the sliding window strategy for the inter-human relation module (for the analysis of the maximum person number N , please see the ablation study). The sliding window divides all persons in the image into multiple N -person groups, and then we apply the inter-human relation module within each group.

4 Experiments

4.1 Datasets and Evaluation Metrics

CrowdPose [Li *et al.*, 2019] is one of the most commonly used datasets in multi-person pose estimation task. The images in this dataset are heavily crowded and challenging. It contains 10K, 2K and 8K images for training set, validation set and testing set, respectively. Following [Cheng *et al.*, 2020; Geng *et al.*, 2021], we train our models on the training set and validation set, and evaluate on the testing set.

OCHuman [Zhang *et al.*, 2019] is a recently proposed benchmark to examine the limitations of MPPE in highly occluded scenarios. It consists of 4731 images, including 2500 images for validation and 2231 images for testing. Following [Qiu *et al.*, 2020], we train models on the validation set and report the performance on the testing set.

COCO [Lin *et al.*, 2014] is one popular benchmark for keypoint detection. Although it contains few crowded scenes, we use it to validate the generation ability of our method for non-crowded scenes. We train the models on the training set and report the results on the validation set.

Evaluation Metrics. We follow the standard evaluation procedure and use OKS-based metrics for pose estimation [Lin *et al.*, 2014]. We report average precision (AP), average recall (AR) with different thresholds. Following [Li *et al.*, 2019], on CrowdPose, we also report AP_{easy} , AP_{medium} and AP_{hard} , which refers to the performance under different crowd index.

4.2 Implementation Details

We use Adam optimizer and the cosine annealing learning rate decay from $1e-4$ to $1e-5$. We also adopt the coordinate decoding strategy in [Zhang *et al.*, 2020] to reduce the quantisation error when decoding from downscaled heatmaps. For the training batch size, we choose 64 for I^2R -Net at 256×192 resolution, and 32 at 384×288 resolution. The length of each token channel d is 96. The number of transformer layers $M = 6$ in the vanilla I^2R -Net, and $M = 2$ for other models. Each experiment takes $8 \times 32G$ -V100 GPUs for training.

Method	AP	AP ₅₀	AP ₇₅	AR	AR ₅₀	AR ₇₅	AP _{easy}	AP _{medium}	AP _{hard}
Bottom-up									
OpenPose [Cao <i>et al.</i> , 2019]	-	-	-	-	-	-	62.7	58.7	32.3
HigherHRNet [Cheng <i>et al.</i> , 2020]	65.9	86.4	70.6	-	-	-	73.3	66.5	57.9
HigherHRNet Multi-scale [Cheng <i>et al.</i> , 2020]	67.6	87.4	72.6	-	-	-	75.8	68.1	58.9
SPM [Nie <i>et al.</i> , 2019]	63.7	85.9	68.7	-	-	-	70.3	64.5	55.7
DEKR [Geng <i>et al.</i> , 2021]	65.7	85.7	70.4	-	-	-	73.0	66.4	57.5
DEKR Multi-scale [Geng <i>et al.</i> , 2021]	67.0	85.4	72.4	-	-	-	75.5	68.0	56.9
PINet [Wang <i>et al.</i> , 2021a]	68.9	88.7	74.7	-	-	-	75.4	69.6	61.5
PINet Multi-scale [Wang <i>et al.</i> , 2021a]	69.9	89.1	75.6	-	-	-	76.4	70.5	62.2
Top-down									
Mask R-CNN [He <i>et al.</i> , 2017]	57.2	83.5	60.3	65.9	89.5	69.4	69.4	57.9	45.8
AlphaPose [Fang <i>et al.</i> , 2017]	61.0	81.3	66.0	67.6	86.7	71.8	71.2	61.4	51.1
Simple baseline [Xiao <i>et al.</i> , 2018]	60.8	81.4	65.7	67.3	86.3	71.8	71.4	61.2	51.2
CrowdPose [Li <i>et al.</i> , 2019]	66.0	84.2	71.5	72.7	89.5	77.5	75.5	66.3	57.4
OPEC-Net [Qiu <i>et al.</i> , 2020]	70.6	86.8	75.6	-	-	-	-	-	-
HRNet [Sun <i>et al.</i> , 2019]	71.3	91.1	77.5	-	-	-	80.5	71.4	62.5
HRNet† [Sun <i>et al.</i> , 2019]	72.8	92.1	78.7	-	-	-	81.3	73.3	65.5
TransPose-H [Yang <i>et al.</i> , 2021]	71.8	91.5	77.8	75.2	92.7	80.4	79.5	72.9	62.2
HRFormer-B [Yuan <i>et al.</i> , 2021]	72.4	91.5	77.9	75.6	92.7	81.0	80.0	73.5	62.4
<i>I</i> ² R-Net (Vanilla version, 1st stage: HRNet)	72.3	92.4	77.9	76.5	93.2	81.9	79.9	73.2	62.8
<i>I</i> ² R-Net (1st stage: TransPose-H)	76.3	93.5	82.2	79.1	94.0	84.4	83.2	77.0	67.4
<i>I</i> ² R-Net (1st stage: HRFormer-B)	77.4	93.6	83.3	80.3	94.5	85.5	83.8	78.1	69.3

Table 1: Comparisons with state-of-the-art methods on the CrowdPose dataset. The default input resolution for experiments is 256×192 , while † denotes the input resolution is 384×288 .

Method	Input size	#Params	FLOPs	AP	AP ₅₀	AP ₇₅	AP _M	AP _L	AR	AR _M	AR _L
HRNet-W32 [Sun <i>et al.</i> , 2019]	256×192	28.5M	7.1G	74.4	90.5	81.9	70.8	81.0	78.9	75.7	85.8
HRNet-W48 [Sun <i>et al.</i> , 2019]	256×192	63.6M	14.6G	75.1	90.6	82.2	71.5	81.8	80.4	76.2	86.4
<i>I</i> ² R-Net (Vanilla version, 1st stage:HRNet-W48-S)	256×192	18.0M	10.2G	75.3	90.2	81.9	71.7	82.4	80.5	76.1	86.8
TransPose-H [Yang <i>et al.</i> , 2021]	256×192	17.5M	21.8G	75.8	90.1	82.1	71.9	82.8	80.8	76.4	87.2
<i>I</i> ² R-Net (1st stage:TransPose-H)	256×192	18.0(+0.5)M	22.7(+0.9)G	75.8	90.4	82.1	72.0	82.9	80.9	76.6	87.3
HRFormer-B* [Yuan <i>et al.</i> , 2021]	256×192	43.2M	12.2G	75.6	90.8	82.8	71.7	82.6	80.8	76.5	87.2
<i>I</i> ² R-Net (1st stage:HRFormer-B*)	256×192	43.7(+0.5)M	12.8(+0.6)G	76.4	90.8	83.2	72.3	83.7	81.4	76.9	88.1
HRFormer-B* [Yuan <i>et al.</i> , 2021]	384×288	43.2M	26.8G	77.1	90.9	83.5	69.8	80.5	82.0	77.7	88.3
<i>I</i> ² R-Net (1st stage:HRFormer-B*)	384×288	43.7(+0.5)M	29.6(+2.8)G	77.3	91.0	83.6	73.0	84.5	82.1	77.7	88.6

Table 2: Results on the COCO pose estimation *validation* set. The results of methods marked with * are obtained by running the official code and model. Following TransPose, HRNet-W48-S only uses early 3 stages of HRNet-W48.

Method	AP	AP ₅₀	AP ₇₅
Bottom-up			
SPM [Nie <i>et al.</i> , 2019]	47.6	67.5	53.2
HigherHRNet [Cheng <i>et al.</i> , 2020]	27.7	66.9	15.9
DEKR [Geng <i>et al.</i> , 2021]	52.2	69.9	56.6
PINet [Wang <i>et al.</i> , 2021a]	59.8	74.9	65.9
Top-down			
Mask R-CNN [He <i>et al.</i> , 2017]	20.2	33.2	24.5
SimplePose [Xiao <i>et al.</i> , 2018]	24.1	37.4	26.8
CrowdPose [Li <i>et al.</i> , 2019]	27.5	40.8	29.9
OPEC-Net [Qiu <i>et al.</i> , 2020]	29.1	41.3	31.4
HRFormer-B [Yuan <i>et al.</i> , 2021]	62.1	81.4	67.1
TransPose-H [Yang <i>et al.</i> , 2021]	62.3	82.7	67.1
<i>I</i> ² R-Net (Vanilla version, 1st stage: HRNet)	64.3	85.0	69.2
<i>I</i> ² R-Net (1st stage: HRFormer-B)	66.5	83.8	71.4
<i>I</i> ² R-Net (1st stage:TransPose-H)	67.8	85.0	72.8

Table 3: Comparisons with state-of-the-art methods on the OCHuman *testing* set after training on OCHuman *validation* set.

4.3 Compare with the State-of-the-art Methods

Results on the CrowdPose testing set. As shown in Table 1, thanks to our Inter-human relation module, the *I*²**R-Net** (1st stage: TransPose-H) achieves 76.3% AP, which outperforms TransPose-H. Moreover, the *I*²**R-Net** (1st stage: HRFormer-B) achieves 77.4% AP at 256×192 resolution, even surpassing the performance of HRNet† at resolution of 384×288 . This demonstrates that our method is very suitable for crowded scenes, and performs better than other methods even without bells and whistles.

Results on the OCHuman testing set. On OCHuman dataset, our method also obtains superior performance. The results are shown in Table 3. The vanilla *I*²**R-Net** achieves 64.3% AP on OCHuman test set, which is comparable with state-of-the-art methods. Furthermore, it is clear in Table 3 that, equipped with the Inter-Human Relation Module, HRFormer-B and TransPose-H are significantly improved, reaching 66.5% and 67.8% AP, respectively.



Figure 4: **Visualization Comparisons.** From left to right, each example shows the results of TransPose, HRFormer, and I^2R -Net. The incorrect regions are marked with white circles.



Figure 5: **Inter-Human Attention Visualization.** The incorrect region (marked in the red box) is corrected with the aid of higher response regions, which are indicated by the orange arrows.

Results on the COCO pose estimation validation set. As shown in Table 2, I^2R -Net achieves 77.3% AP, which is competitive against state-of-the-art methods. The performance

gain on COCO is not that large compared with OCHuman and CrowdPose. This is because that COCO contains around 48% single-person images. Hence, our method for multi-person scenarios is not able to give full play. But this result demonstrates that our method also works well in non-crowded scenes.

4.4 Ablation Studies

Effectiveness of Inter-Human Relation Module. On the CrowdPose and OCHuman, I^2R -Net outperforms all existing methods at the same resolution. On COCO datasets, I^2R -Net also achieves competitive performance against state-of-the-art methods, especially outperforming on AP_M and AP_L . As shown in Table 1, under our two-stage framework, TransPose can significantly surpass HRNet[†] with Inter-Human Relation

Module even at lower resolution. This demonstrates that our Inter-Human Relation Module can bring significant performance improvements.

As is discussed in Section 3.3, the relation modeling capability of inter-human relation module is not limited to cross-instance, but can also relates parts in-instance. This can be evidenced in Table 1, 2 and 3, where the vanilla Inter-Human Relation (without intra- stage, and extract per-person feature using a backbone) achieves comparable performance with state-of-the-art methods.

Instance patch number. An important factor of I^2R -Net is the number of instances fed into the Inter-human relation module. We compare the results of different instance numbers on CrowdPose in Table 4a. We can observe that the AP value tends to be saturated when N is 4 and reaches the highest when N is 6. A higher or lower number of patches will cause a drop in AP value. This is because too few patches may lose information between persons, while too many patches introduce unnecessary cross-human correlations.

Effectiveness of Intermediate-supervision. In our framework, we use MSE loss in the first stage as intermediate-supervision (denoted as IS). As shown in the second and last rows in Table 4b, the intermediate-supervision improves the performance.

Position encoding. We also try to add position encoding into I^2R -Net. As shown in Table 4b, neither using sinusoidal mode nor encoding the global position as a positional encoding can improve the results. One possible reason is that the Intra-Human Relation Module has learned the local correlations between parts of each person, and the global interactions between instances are position-independent.

Person selection mechanism. We compare two strategies for person selection. One is the nearest neighbor strategy, which randomly selects a target person and selects $N - 1$ closest neighbors; the other is to completely randomly select N persons. As shown in Table 4c, there are no significantly difference between them, which means the multi-person interaction is position-free.

Token size setting. In Table 4d, the setting of $16 \times 12 \times N$ achieves a better result. It shows that our method can learn enough inter-human relations at low resolution, but $1 \times 1 \times N$ leads to low precision due to the lack of spatial information.

4.5 Qualitative Analysis

In Figure 4, we visualize our results compared with two competitive state-of-the-art methods (TransPose and HRFormer). The proposed method is able to handle the heavily crowded cases and localize the key points correctly. Even though in occluded scenes, we can also infer plausible poses.

By visualizing the attention maps between persons, the performance gain can be justified. As shown in Figure 5, the incorrect localized keypoints are corrected by highly related regions of the other person.

4.6 Model Size and Computation Cost

As shown in Table 2, Inter-Human Relation Module is lightweight yet effective. In the two-stage framework,

Patch	AP	AR
1	74.6	77.5
2	75.7	78.7
4	76.1	79.0
6	76.3	79.1
8	76.2	79.1

(a) Instance patch number.

Strategy	AP	AR
Nearest	76.0	79.0
Random	76.0	78.9

(c) Person selection strategy.

Model	AP
Base	66.2
Base+IS	66.5
Base+pe _{sine}	66.2
Base+pe _{global}	65.3
Base+pe _{global} +IS	65.7

(b) Effectiveness of components in I^2R -Net.

Token Size	AP
$16 \times 12 \times N$	66.5
$32 \times 24 \times N$	66.4
$1 \times 1 \times N$	66.0

(d) Token size setting.

Table 4: Ablation study of our I^2R -Net with different settings.

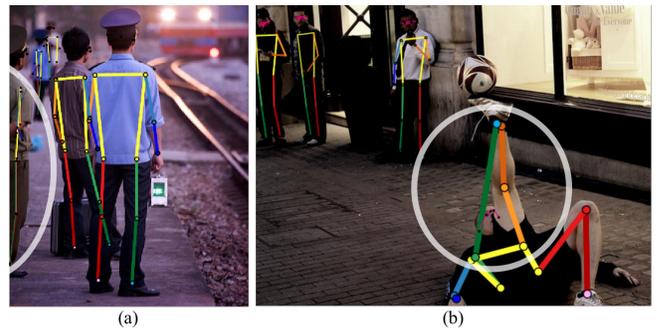


Figure 6: **Failure cases.** The incorrect regions are marked with white circles.

for instance, HRFormer with Inter-Human Relation Module merely increased the parameters by 0.5M and FLOPS by 0.6G. This demonstrates that our Inter-Human Relation Module cost a few and acceptable extra computations but achieves significant improvement for multi-person estimation.

4.7 Failure Cases

Although our approach outperforms the state-of-the-art methods, we still suffer from several extremely challenging cases. Like previous work, our method may fail in cases where only part of the person is visible (Figure 6(a)) or the inter-human relationships are insignificantly correlated (Figure 6(b)). All these cases with poor relations may lead to failure predictions.

5 Concluding Remarks

This work investigates and evidences that collaboratively estimating poses via modeling multi-person correlation promotes pose estimation. We propose a flexible and efficient framework to integrate intra-human and inter-human information, leading to superior results than state-of-the-art methods. Inspired by the idea of this work, promising future work is to extend the inter-human relation from the spatial dimension to the temporal dimension. Another interesting direction is to explore the performance gain via correlating the human-object relationships in various vision tasks.

Acknowledgements

This work was supported in part by NSFC (No. 62072382) and Yango Charitable Foundation.

References

- [Bello *et al.*, 2019] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *CVPR*, 2019.
- [Cao *et al.*, 2019] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 2019.
- [Chen *et al.*, 2018] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, 2018.
- [Cheng *et al.*, 2020] 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 *CVPR*, 2020.
- [Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [Fang *et al.*, 2017] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2017.
- [Geng *et al.*, 2021] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *CVPR*, 2021.
- [He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.
- [Jin *et al.*, 2020] Sheng Jin, Wentao Liu, Enze Xie, Wenhui Wang, Chen Qian, Wanli Ouyang, and Ping Luo. Differentiable hierarchical graph grouping for multi-person pose estimation. In *ECCV*, 2020.
- [Li *et al.*, 2019] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*, 2019.
- [Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.
- [Mihai *et al.*, 2021] Fieraru Mihai, Zanfir Mihai, Alexandru Szenté Teodor, Gabriel Bazavan Eduard, Olaru Vlad, and Cristian Sminchisescu. Remips: Physically consistent 3d reconstruction of multiple interacting people under weak supervision. *NeurIPS*, 2021.
- [Newell *et al.*, 2016a] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *NeurIPS*, 2016.
- [Newell *et al.*, 2016b] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.
- [Nie *et al.*, 2019] Xuecheng Nie, Jiashi Feng, Jianfeng Zhang, and Shuicheng Yan. Single-stage multi-person pose machines. In *CVPR*, 2019.
- [Papandreou *et al.*, 2018] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *ECCV*, 2018.
- [Parmar *et al.*, 2018] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *ICML*, 2018.
- [Qiu *et al.*, 2020] 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 *ECCV*, 2020.
- [Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [Sun *et al.*, 2019] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.
- [Touvron *et al.*, 2021] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021.
- [Wang *et al.*, 2021a] Dongkai Wang, Shiliang Zhang, and Gang Hua. Robust pose estimation in crowded scenes with direct pose-level inference. *NeurIPS*, 34, 2021.
- [Wang *et al.*, 2021b] Jiashun Wang, Huazhe Xu, Medhini Narasimhan, and Xiaolong Wang. Multi-person 3d motion prediction with multi-range transformers. *NeurIPS*, 2021.
- [Xiao *et al.*, 2018] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018.
- [Yang *et al.*, 2021] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Keypoint localization via transformer. In *ICCV*, 2021.
- [Yuan *et al.*, 2021] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution transformer for dense prediction. *NeurIPS*, 2021.
- [Zhang *et al.*, 2019] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *CVPR*, 2019.
- [Zhang *et al.*, 2020] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *CVPR*, 2020.