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

Yan Dai, Xuanhan Wang, Lianli Gao,
Jingkuan Song, Heng Tao Shen

Dai, Y., Wang, X., Gao, L., Song, J., & Shen, H. T. (2021). RSGNet: Relation based Skeleton Graph Network for Crowded Scenes Pose Estimation. In *35th AAAI Conference on Artificial Intelligence.*

Outline

Task Definition

Challenges and Motivations

Proposed Approach

Experiments

Conclusion

Task Definition

Crowded Scenes Pose Estimation

Task? → Localize the anatomical joints of each person from a given image.

Crowded Scenes? → Complex real-world scenes with highly-overlapped people, severe occlusions and diverse postures.

Challenges and Motivations

Existing challenges applying top-down pipelines

Fig.1. Multi-joints in one bounding box.

Challenge 1: Since a generated bounding box contains both target joints and interference joints, an identical joint is assigned with different labels and missing joints cannot be restored.

→ Encourage all joints in one bounding box to be active.

Challenge 2: A joint-to-joint relation modeling method and the human body structure priors are needed for interference removal.

→ Enforce such priors during the joints inference.

Challenges and Motivations

Our motivations:

1) how to design an effective pipeline for ***crowded scenes*** pose Estimation.

2) how to equip this pipeline with the ability of ***relation modeling*** for interference resolving.

A multi-joints representation with relation modeling.

Proposed Approach

Framework of RSGNet

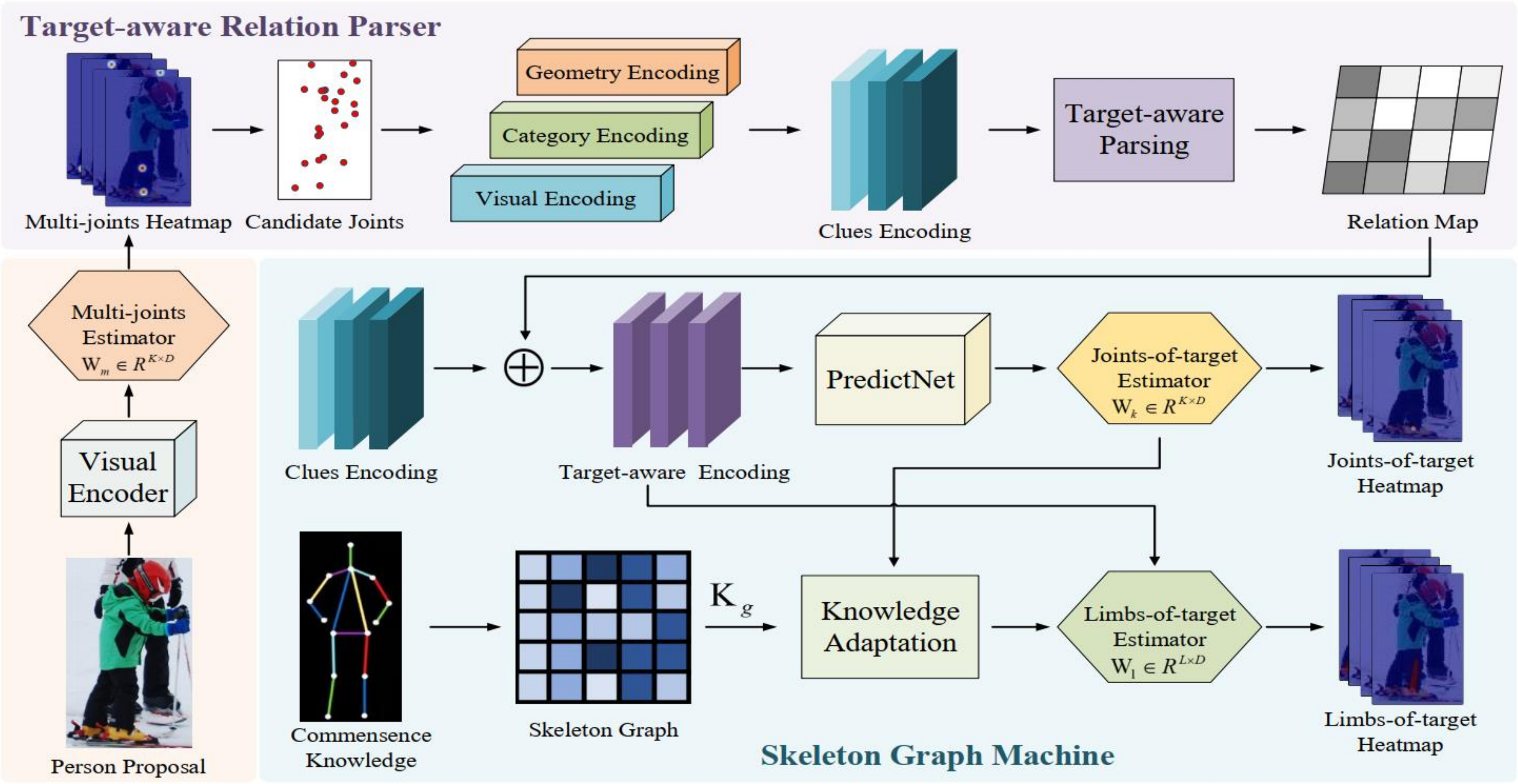

Fig.2. The framework of our proposed RSGNet, which consists of a CNN based visual encoder, a target-aware relation parser, and a skeleton graph machine.

Proposed Approach

Target-aware Relation Parser (TRP)

Step 1: Generate candidate joints from the obtained multi-joints heatmap, and encode the information of joint semantic, joint location and visual appearance to form a clues encoding.

Step 2: Construct a joint-to-joint relation map through the target-aware parsing for interference resolving, and generate a target-aware encoding.

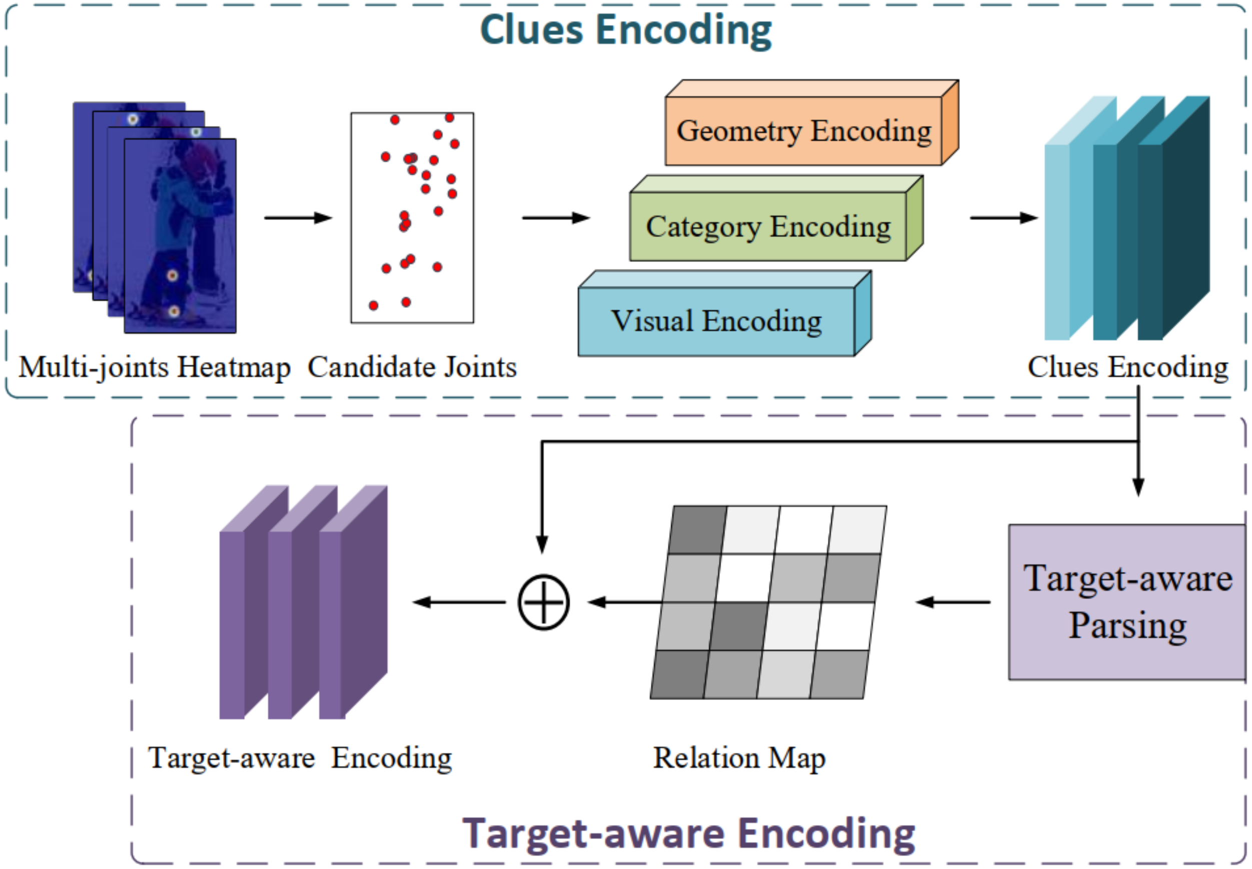

Fig.3. Illustration of proposed TRP module.

Proposed Approach

Skeleton Graph Machine (SGM)

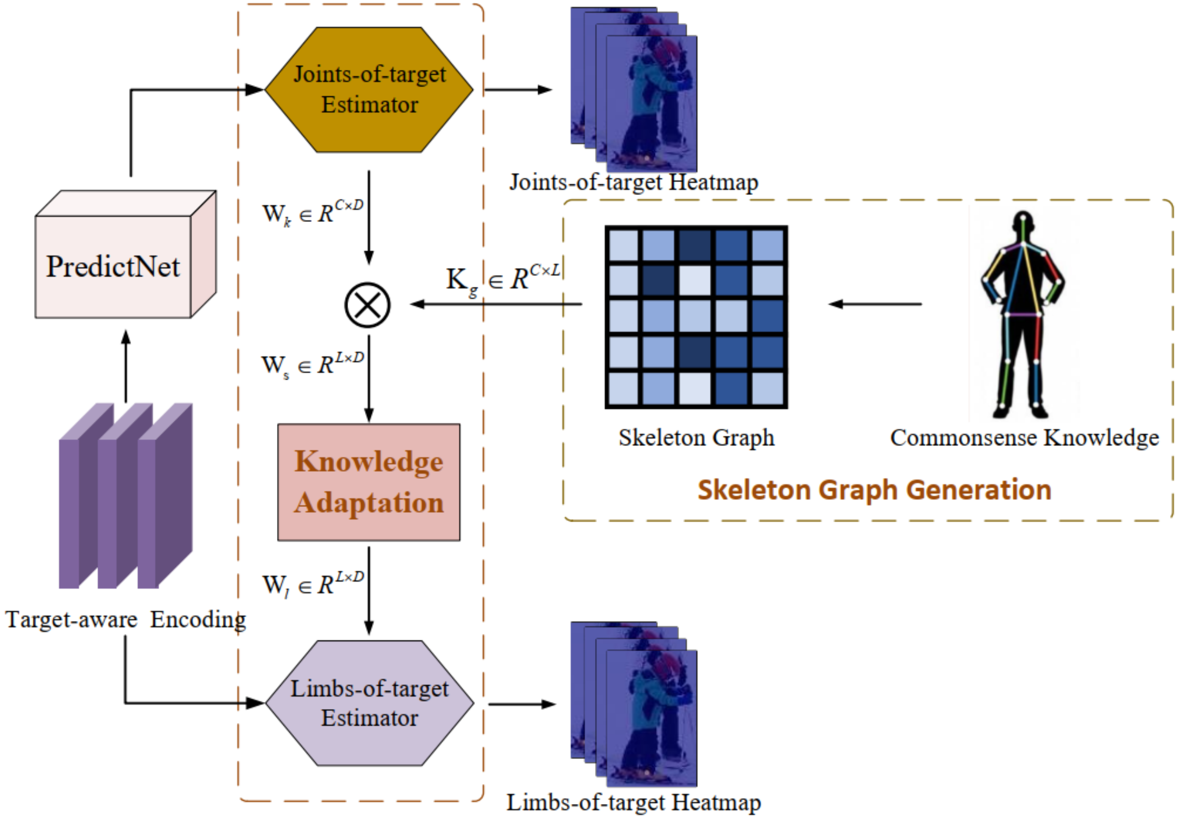

Fig.4. Illustration of proposed SGM module.

Step 1: Create a skeleton-based graph, and provides relation among joints and limbs.

Step 2: Transform the parameters of joints estimator into parameters of limbs estimator through the knowledge adaptation, and therefore, the joints estimation results, can be constrained by human body structure priors.

Experiments

Dataset

CrowdPose

This dataset contains **12K**, and **8K** images for training and testing, respectively. It has approximately **80k** human annotations totally, and **14** human joints annotations for each human instance.

MSCOCO

This dataset contains over **60K** images and **250K** person instances annotated with **17** human joints. Moreover, it is divided into **57K** , **5K** and **20K** images for training, validation and testing, respectively.

Evaluation Metric

mAP: the mean of AP scores at a number of object keypoints similarity (OKS) ranging from 0.5 to 0.95.

Experiments

Ablation Studies

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**

Tab.1. Investigating the effect of proposed modules.

Input resolution: 256×192

Different models:

- HRNet-W32(baseline)
- HRNet-W32 with TRP only
- HRNet-W32 with TRP and SGM(Our RSGNet)

Experiments

Comparison Results

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 (Ke Sun and Wang 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 (Ke Sun and Wang 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 (Ke Sun and Wang 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**

Tab.2. 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
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 (Ke Sun and Wang 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 (Ke Sun and Wang 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 (Ke Sun and Wang 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 (Ke Sun and Wang 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**

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

11

Experiments

Quantitative Analysis

Fig.5. Qualitative results comparison on CrowdPose *test* set.

Conclusion

Our contributions:

1. Cast the crowded problem of pose estimation as an **interference resolution problem**.
2. Design a ***target-aware relation parser (TRP)*** for interference removal.
3. Propose a ***skeleton graph machine (SGM)*** to enforce the constraint of human body.
4. Significantly **outperforms** current state-of-the-art pose estimation methods, especially on the CrowdPose dataset.

Thank you!

The code is releasd on GitHub:

https://github.com/vikki-dai/RSGNet

If you have any questions, please e-mail us at:

yandai1019@gmail.com

wxuanhan@hotmail.com