A Multi-level Network for Human Pose Estimation

Zhanpeng Shao[1] , Peng Liu[1], Y.F. LI[2], Jianyu Yang[3], and Xiaolong Zhou[4]

[1] College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China
zpshao@zjut.edu.cn

[2] Department of Mechanical Engineering, City University of Hong Kong, Hong Kong, China

[3] College of Electrical and Information Engineering, Quzhou University, Quzhou, China

IEEE ICRA 2021

浙江工业大学 Zhejiang University of Technology

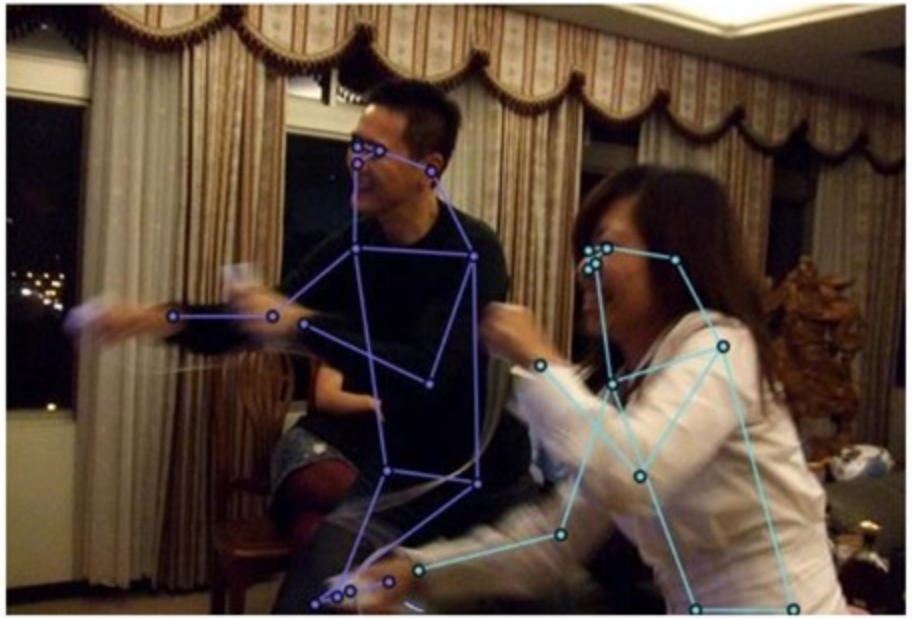
PART ONE

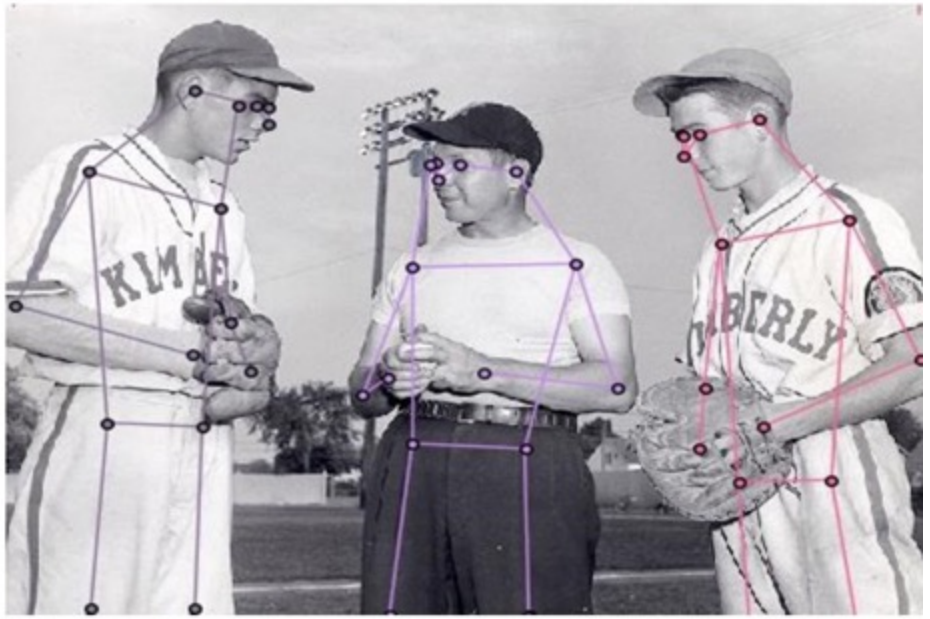
Introduction

Human Pose Estimation

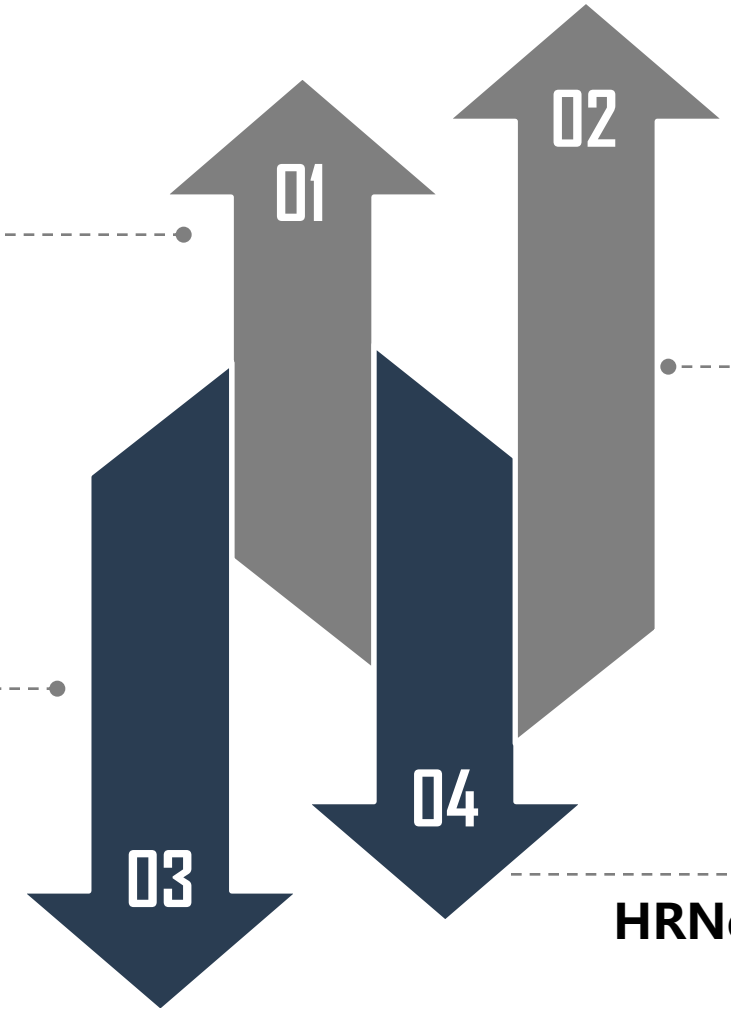
locate and recognize the keypoints for all human bodies in an image.

- different scale
- occluded keypoints
- Crowded backgrounds
- ⋮





Excellent solution of HPE



Open Pose

Pifpaf

CPN (Cascaded Pyramid Network)

HRNet (High-Resolution Network)

Why MLPE is proposed

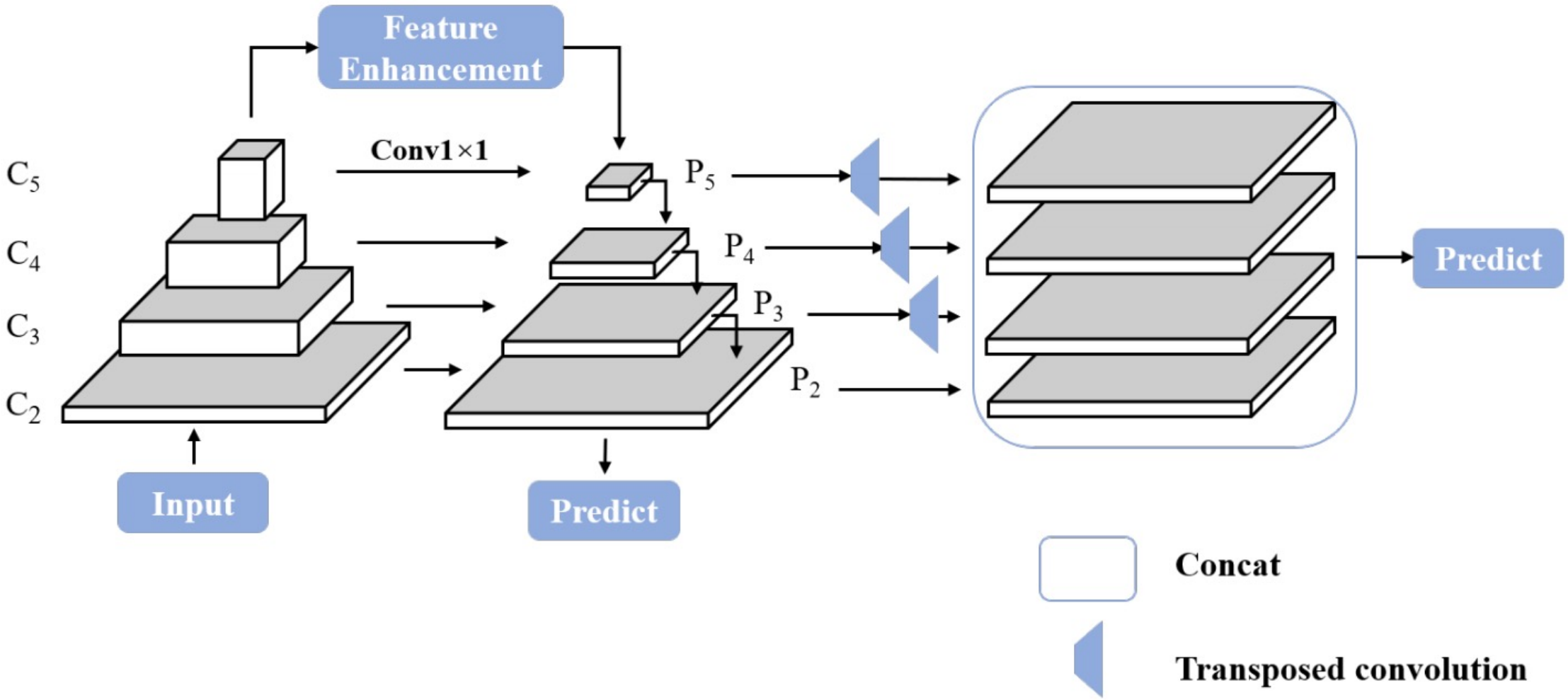
The feature extraction process loses more information	Backbone network uses ResNeSt	MLPE
High-level feature transmission loses semantic information	Feature enhancement strategy	
Spatial resolution information is not fully utilized	High-resolution fine-tuning network	

02

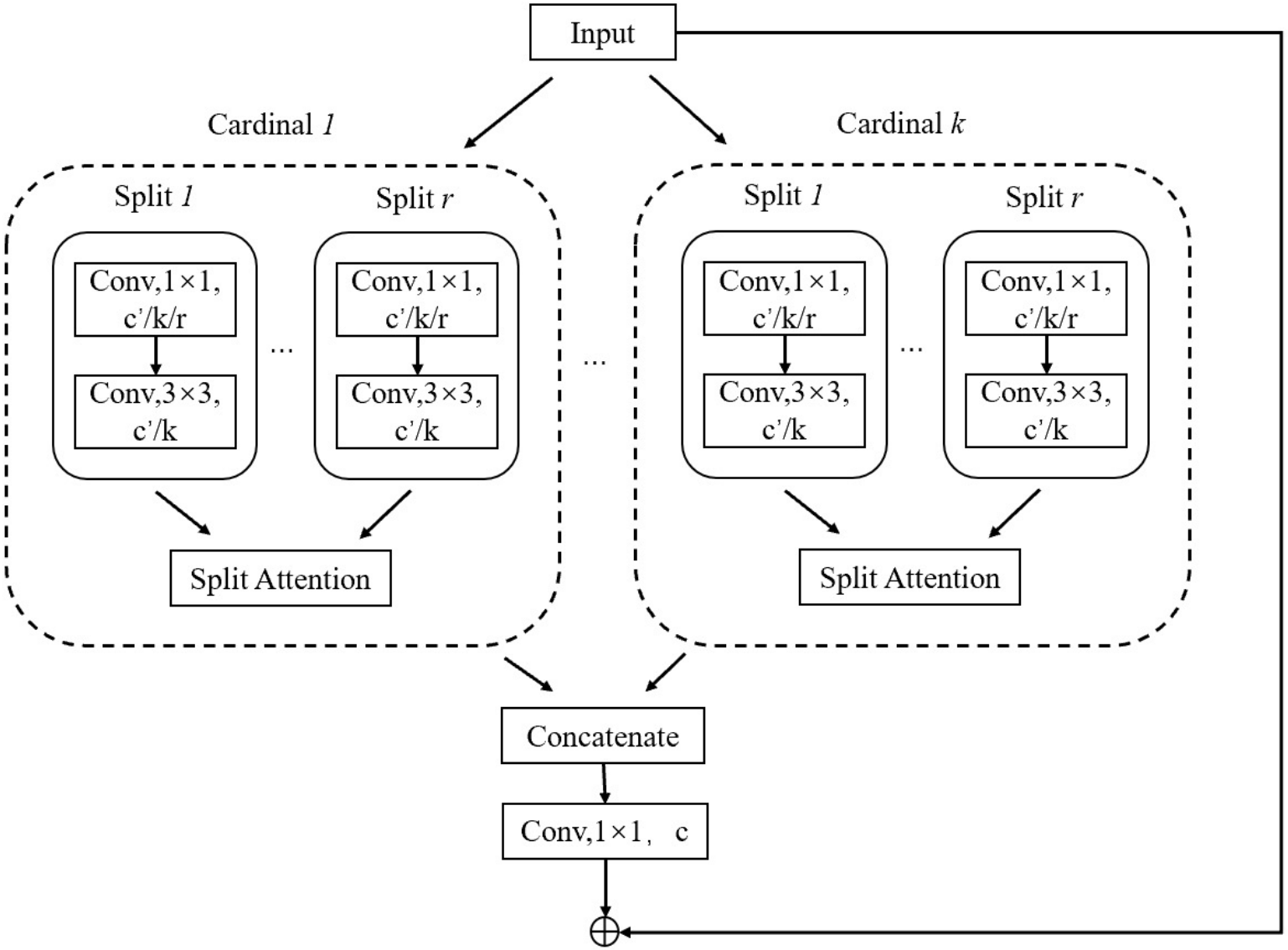
PART TWO

Approach

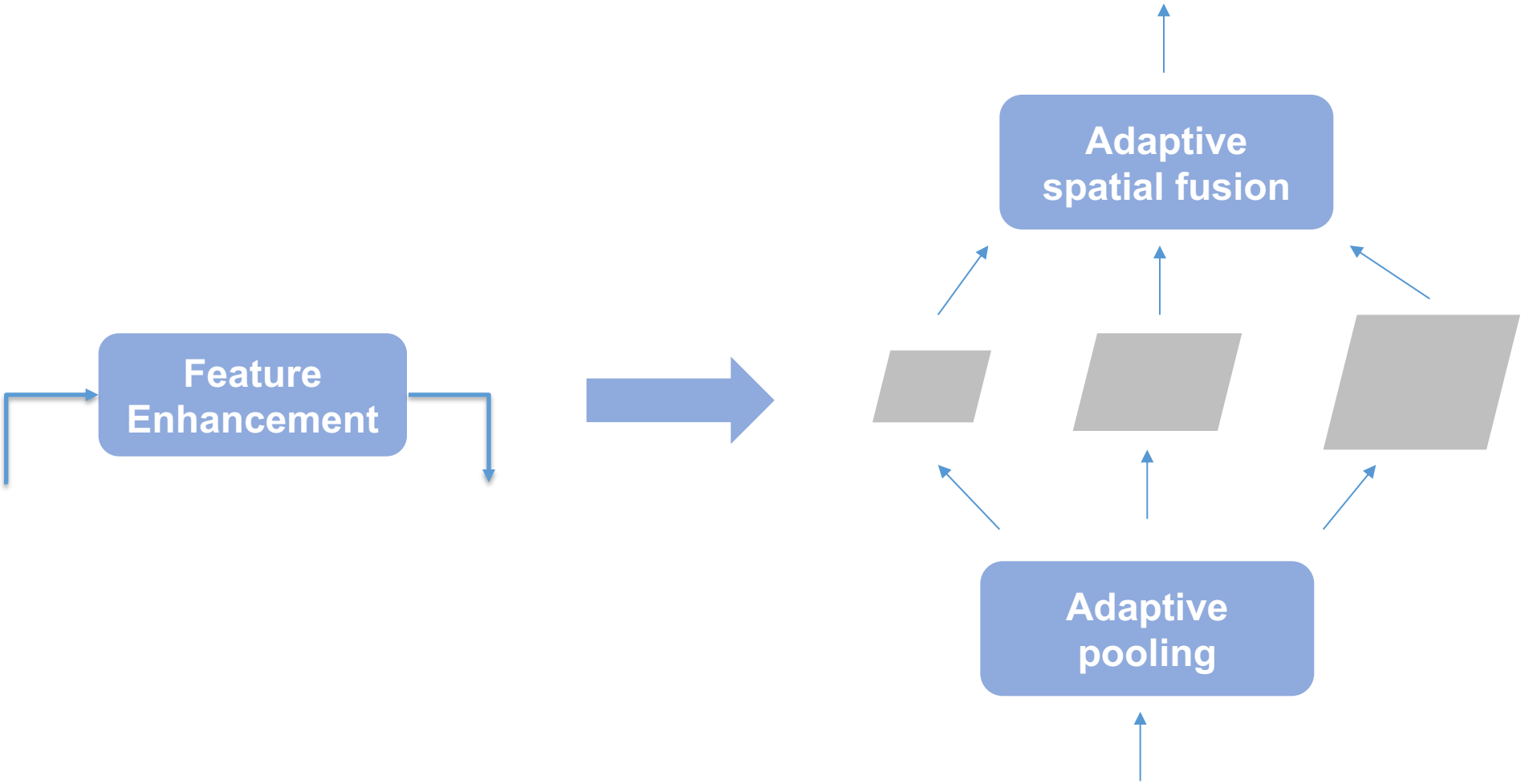
MLPE framework



ResNeSt module：



Feature Enhancement



PART THREE

Experiment

Dataset and Evaluation criteria

MS COCO Dataset：

The number of person instances exceeds 250K and each person is annotated with 17 keypoints.

COCO train2017: 57K images、 150K individual instances

COCO val2017: 5000 images

COCO test-dev: 20K images

AP(Average Precision)

$$AP = \frac{\sum_p (OKS_p > T)}{\sum_p 1} \qquad OKS_p = \frac{\sum_i \exp(-d_{pi}^2 / 2S_p^2\sigma_i^2)\delta(\theta_i > 0)}{\sum_i \delta(\theta_i > 0)}$$

Ablation experiments of MLPE

TABLE I

ABLATION STUDY OF OUR METHOD ON COCO VAL2017 DATASET. BI AND DECONV REPRESENT TWO UPSAMPLING METHODS RESPECTIVELY, BI REFERS TO BILINEAR INTERPOLATION, AND DECONV REFERS TO UPSAMPLING USING TRANSPOSED CONVOLUTION.

Method	Backbone	Input size	Feature enhancement	Receptive field	Upsample	AP	AP^M	AP^L
a	ResNeSt-50	256×192	×	3	BI	70.1	66.5	76.7
b	ResNeSt-50	256×192	√	3	BI	70.2	66.6	77.1
c	ResNeSt-50	256×192	√	3	Deconv	70.4	66.9	77.1
d	ResNeSt-50	256×192	√	1	Deconv	70.2	66.7	76.8
e	ResNeSt-50	384×288	√	3	Deconv	72.4	68.1	79.9
f	ResNeSt-101	384×288	√	3	Deconv	73.7	69.7	81.0

- Feature enhancement strategy
- How to get high resolution features
- Data processing

Comparison with State-of-Art Methods

TABLE II

COMPARISONS ON COCO TEST-DEV DATASET. * MEANS THAT THE METHOD INVOLVES EXTRA DATA FOR TRAINING.

Method	Backbone	Input Size	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Bottom-up								
OpenPose [9]	-	-	61.8	84.9	67.5	57.1	68.2	66.5
PersonLab [32]	ResNet-152	1401×1401	68.7	89.0	75.4	64.1	75.5	75.4
MultiPoseNet [33]	-	480×480	69.6	86.3	76.6	65.0	76.3	73.5
HigherHRNet [35]	HRNet-W48	640×640	70.5	89.3	77.2	66.6	75.8	74.9
Top-down								
Mask R-CNN [37]	ResNet-FPN	-	63.1	87.3	68.7	57.8	71.4	-
G-RMI [43]	ResNet-101	353×257	64.9	85.5	71.3	62.3	70.0	69.7
G-RMI* [43]	ResNet-101	353×257	68.5	87.1	75.5	65.8	73.3	73.3
CPN [14]	ResNet-Inception	384×288	72.1	91.4	80.0	68.7	77.2	78.5
RPME [36]	PyraNet	320×256	72.3	89.2	79.1	68.0	78.6	-
Ours	ResNeSt-101	384×288	72.8	90.9	80.5	69.1	79.3	79.2

[9] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” in Proceedings of the IEEE conference on CVPR, 2017, pp. 7291–7299.
[32] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson, and K. Murphy, “Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 269–286.
[33] M. Kocabas, S. Karagoz, and E. Akbas, “Multiposenet: Fast multi□person pose estimation using pose residual network,” in Proceedings of the ECCV, 2018, pp. 417–433.
[35] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang,“Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5386–5395.
[37] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask r-cnn,” in ´ Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.
[43] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bre□gler, and K. Murphy, “Towards accurate multi-person pose estimation in the wild,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4903–4911
[14] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, “Cascaded pyramid network for multi-person pose estimation,” in Proceedings of the IEEE conference on CVPR, 2018, pp. 7103–7112.
[36] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, “Rmpe: Regional multi□person pose estimation,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2334–2343.

Conclusion

Conclusion

A novel human pose estimation method named MLPE, which achieves a trade-off between spatial resolution and context information.

The high resolution fine network based on the transposed convolution is then built to fine-tune those keypoints that have large training losses.

Extensive experimental results demonstrate that our method can achieve promising performance on the challenging MS COCO dataset

Thank you for listening

Email：**zpshao@zjut.edu.cn**
Source code：http://

浙江工业大学
Zhejiang University of Technology