Group Pose: A Simple Baseline for End-to-End Multi-person Pose Estimation (Supplementary Materials)

Huan Liu[1,3*] Qiang Chen[2*] Zichang Tan[2] Jiang-Jiang Liu[2] Jian Wang[2] Xiangbo Su[2]
Xiaolong Li[1,3] Kun Yao[2] Junyu Han[2] Errui Ding[2] Yao Zhao[1,3†] Jingdong Wang[2]
[1]Institute of Information Science, Beijing Jiaotong University [2]Baidu VIS
[3]Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, China

1. More details and analysis on inference speed

We measure different end-to-end frameworks with ResNet-50 backbone on a single NVIDIA A100 GPU. To avoid the speed limitation of I/O and randomness of image augmentation, we pre-load test images with one fixed resolution (*e.g.*, 480×800 or 800×1333) into the GPU memory to remove the time cost of inputs pre-processing. Then, we directly send them into different inference models for a fair and precise comparison. Thus, the reported inference time (Time) in Table 6 is faster than their original papers of PETR [1], QueryPose [2] and ED-Pose [3].

Group Pose is faster than previous end-to-end frameworks with complex decoders. This can be explained by that Group Pose only contains a simple transformer decoder, thus eliminating some extra intermediate processes, *e.g.*, an additional query selection[1] in ED-Pose [3].

2. Ablation on across-instance interactions

Group Pose captures $(K+1)$ across-instance interactions over N queries of the same type, including one instance type and K keypoint types, and the interaction designs are to be explored. With the same basic setting in Section 4.3, the following table includes the relevant ablations:

across-instance interactions	AP	AP_M	AP_L
inst-inst & kpt-kpt	**72.0**	**66.8**	**79.7**
only kpt-kpt	71.4	66.4	79.0
only inst-inst	71.0	65.0	79.0

Results validate that same-type across-instance interactions of both the instance (inst-inst) and keypoint (kpt-kpt) queries are essential in Group Pose. The across-instance interactions in Group Pose bring $+0.6$ and $+1.0$ AP gains over only modeling the keypoint queries and instance queries, suggesting the usefulness of promoting information aggregation of same-type queries, thus improving performance, as analyzed in Section 4.3.

3. Ablation on self-attentions implementations

The proposed group self-attentions introduce two types of self-attentions, including within-instance and across-instance self-attentions. In practice, they are implemented with self-attention modules by calculating multiple attention maps in parallel. We ablate the effects of whether sharing the modules in the following table:

self-attention implementations	share	AP	AP_M	AP_L
group self-attentions	×	**72.0**	**66.8**	**79.7**
group self-attentions	✓	71.5	66.3	79.4

We can observe that sharing modules gives a -0.5 AP drop over the unshared one. This is mainly because the two types of self-attentions in Group Pose are responsible for gathering within-instance and across-instance information, respectively. Thus, it is reasonable that the unshared implementation can achieve a better result.

4. Qualitative results on instance query

Group Pose directly utilizes instance query for classification to identify human instances. For studying what instance query looks at to give final results, we visualize the gradient norm of instance query with respect to each pixel in given images, as shown in Figure 1. The gradient norm reflects the degree of change in final results due to each pixel interference, thus showing which pixels the instance query relies on for classification. The results show that the instance query in Group Pose looks at pixels inside human instances of given images even without human box supervision, thus accurately scoring the predicted poses.

5. Visualization results

We visualize the predicted results on MS COCO in Figure 2 and CrowdPose in Figure 3. As can be observed,

*Equal Contribution. Work done when H. Liu is an intern at Baidu.
†Corresponding author. Email: yzhao@bjtu.edu.cn
[1]Called 'fine human query selection' in their paper.

Group Pose performs well on a wide range of poses, including scale variations, motion blur, pose deformations, occlusion, and crowded scenes. The results demonstrate the effectiveness of our design of Group Pose for end-to-end multi-person pose estimation.

6. Limitation

Group Pose shows good results on benchmark datasets, while there are also some failure cases. We find that Group Pose has difficulties in the situation that only contains a small part (*e.g.,* leg, head) of human instances, resulting in confusing prediction of the unlabeled keypoints, as shown in Figure 4. We will conduct deep studies on this problem in future works.

References

[1] Dahu Shi, Xing Wei, Liangqi Li, Ye Ren, and Wenming Tan. End-to-end multi-person pose estimation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11069–11078, 2022. 1

[2] Yabo Xiao, Kai Su, Xiaojuan Wang, Dongdong Yu, Lei Jin, Mingshu He, and Zehuan Yuan. Querypose: Sparse multi-person pose regression via spatial-aware part-level query. In *Advances in Neural Information Processing Systems*, pages 1–14, 2022. 1

[3] Jie Yang, Ailing Zeng, Shilong Liu, Feng Li, Ruimao Zhang, and Lei Zhang. Explicit box detection unifies end-to-end multi-person pose estimation. In *International Conference on Learning Representations*, pages 1–17, 2023. 1

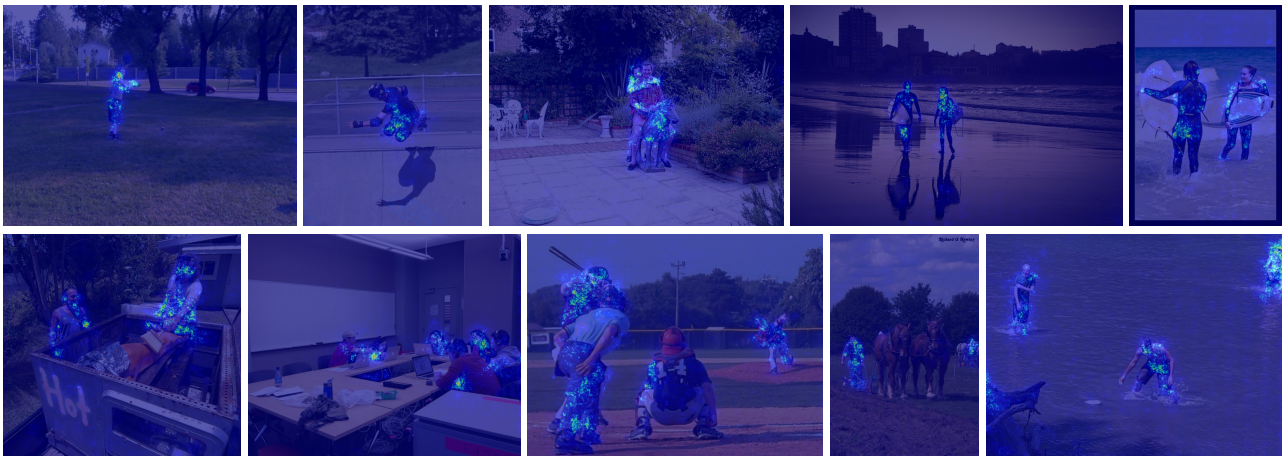

Figure 1: **The gradient norm of instance query** with respect to each pixel in given images. The salient region is visualized by bright color. Best view in zoom in.

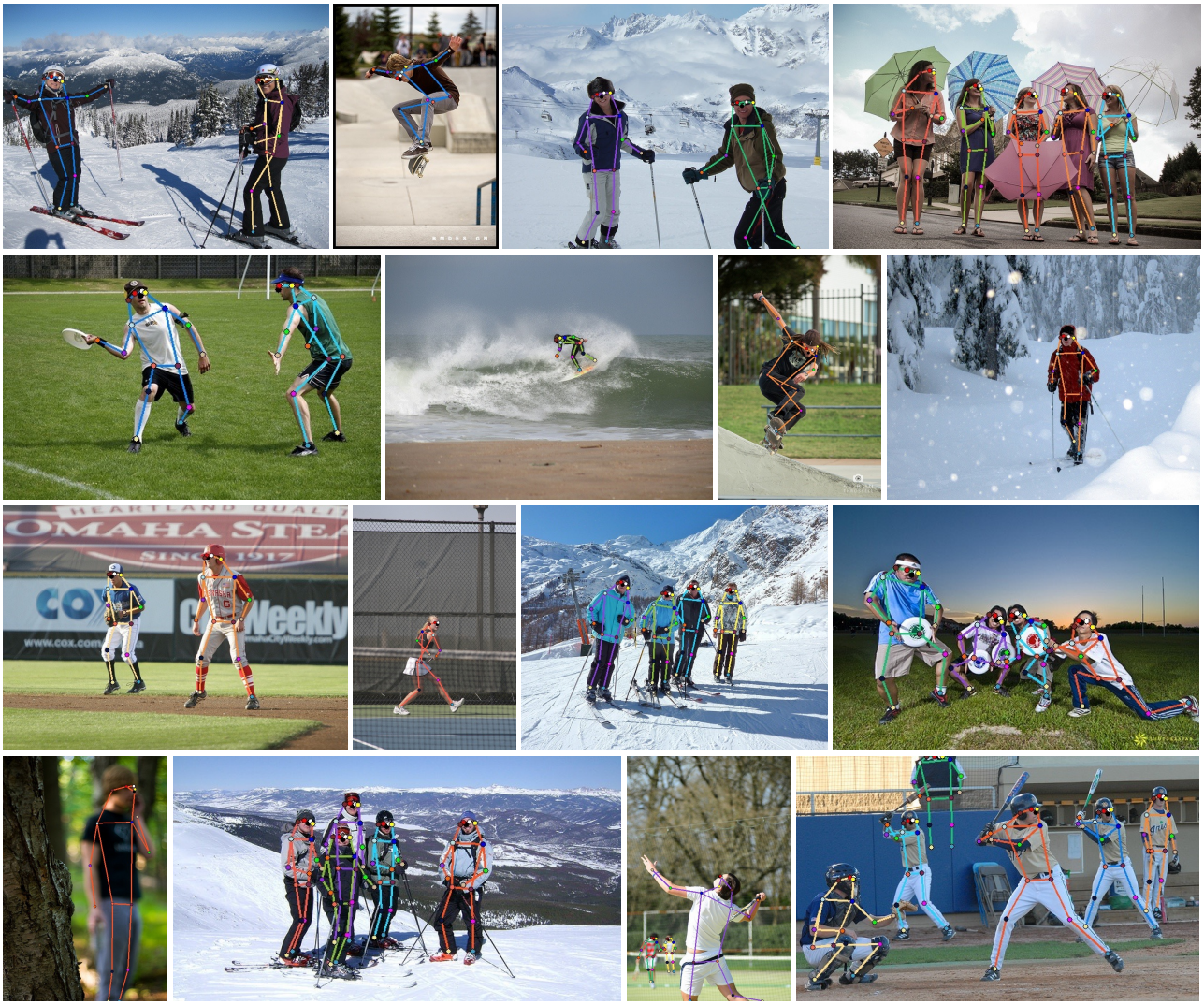

Figure 2: **Visualization results of Group Pose on MS COCO.** Group Pose performs well with scale variations and pose deformations. Best view in color.

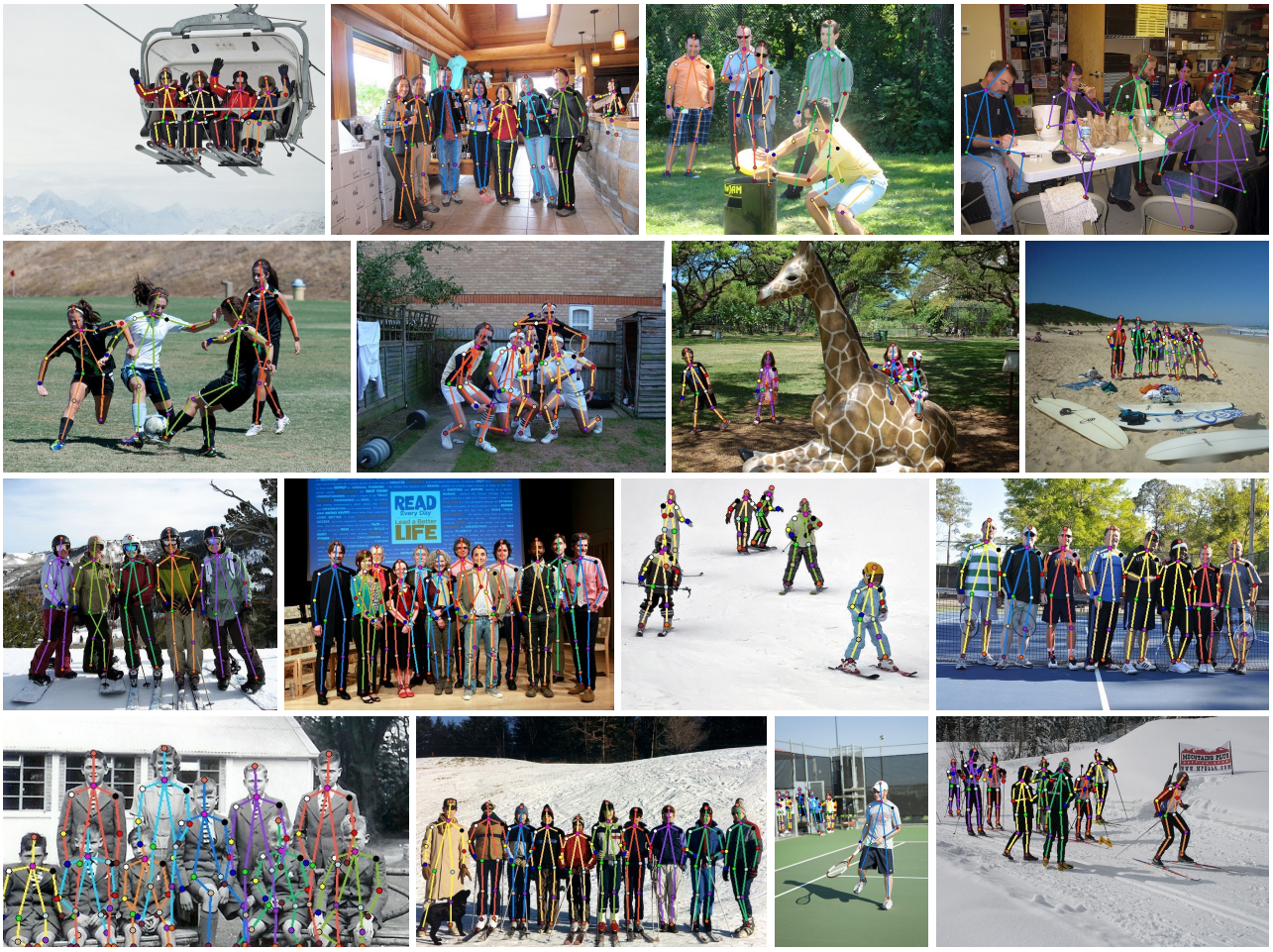

Figure 3: **Visualization results of Group Pose on CrowdPose.** Group Pose is robust for challenging crowded and occluded scenes. Best view in color.

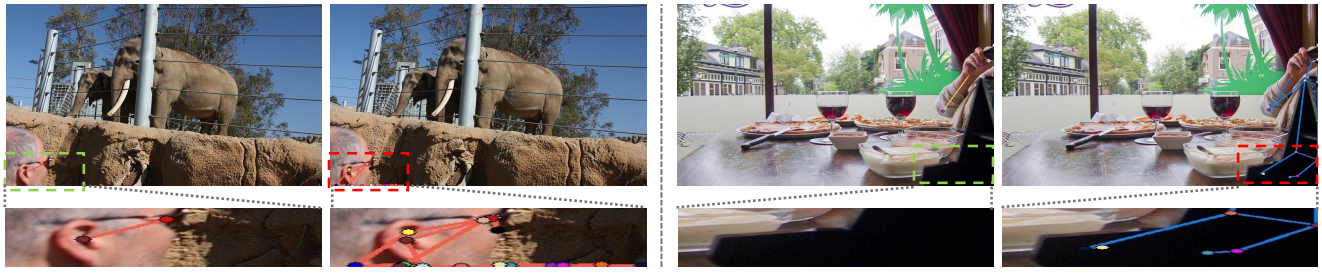

Figure 4: **Visualization results on MS COCO images with small parts of human instance.** Model is based on ResNet-50. The red dashed box indicates difficulties in predicting unlabeled keypoints. Best view in color and zoom in.