Yuanhong ZHONG, Qianfeng XU, Daidi ZHONG, Xun YANG, Shanshan WANG, 2024. FaSRnet: a feature and semantics refinement network for human pose estimation. *Frontiers of Information Technology & Electronic Engineering*, 25(4):513-526. https://doi.org/10.1631/FITEE.2200639

FaSRnet: a feature and semantics refinement network for human pose estimation

Key words: Human pose estimation; Multi-frame refinement; Heatmap and offset estimation; Feature alignment; Multi-person

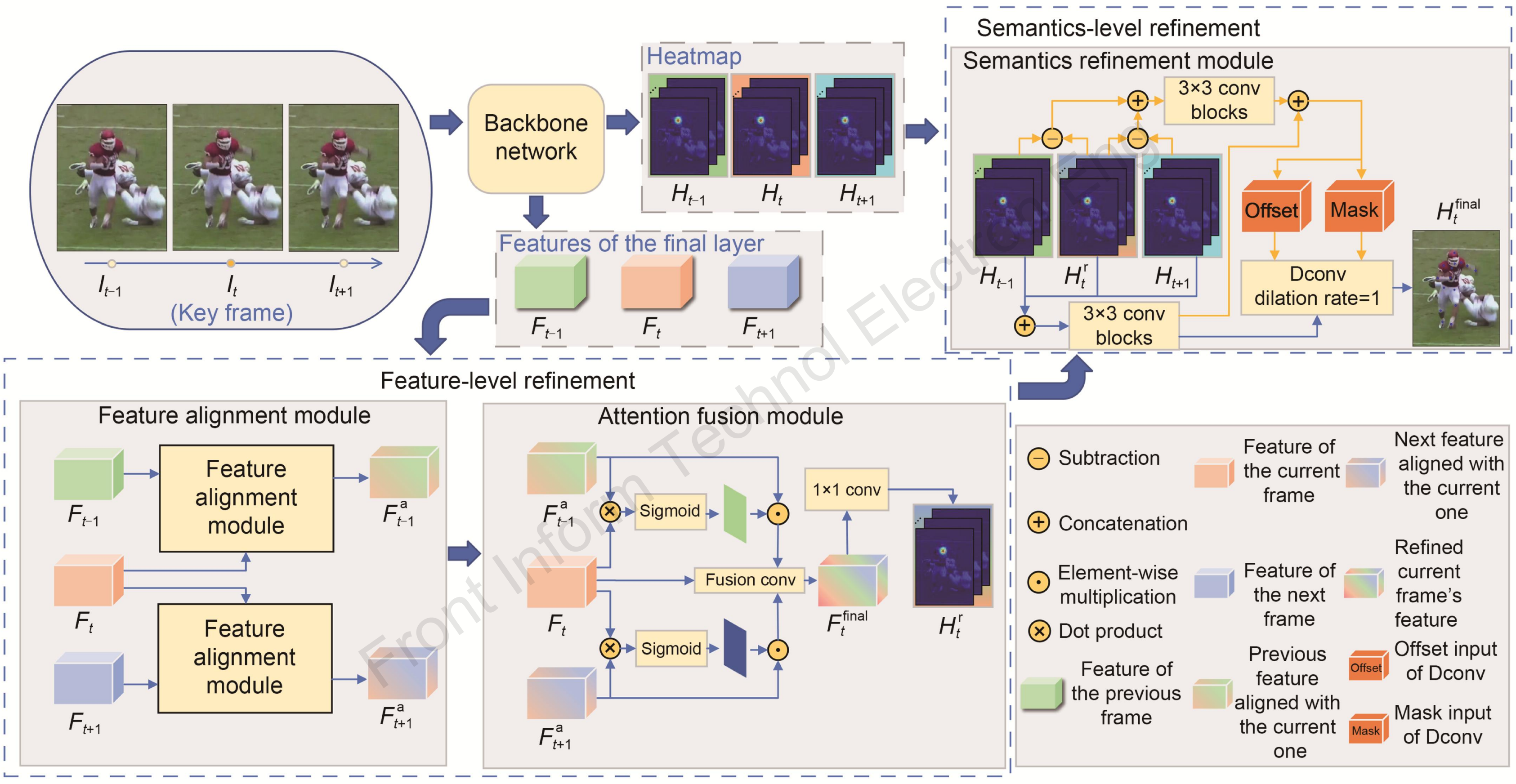
Corresponding author: Yuanhong ZHONG
E-mail: zhongyh@cqu.edu.cn
ORCID: https://orcid.org/0000-0001-5689-1146

Motivation

1. Due to factors such as motion blur, video out-of-focus, and occlusion, multi-frame human pose estimation is a challenging task. Exploiting temporal consistency between consecutive frames is an efficient approach for addressing this issue. Currently, most methods explore temporal consistency through refinements of the final heatmaps.

2. Heatmaps contain the semantics information of key points, and can improve the detection quality to a certain extent. However, the quality of heatmaps is influenced by corresponding features, and directly aggregating heatmaps at the semantics level yields unsatisfactory results. Therefore, we argue that it is necessary to associate and fuse temporal information at the feature level to better address these problems.

Framework



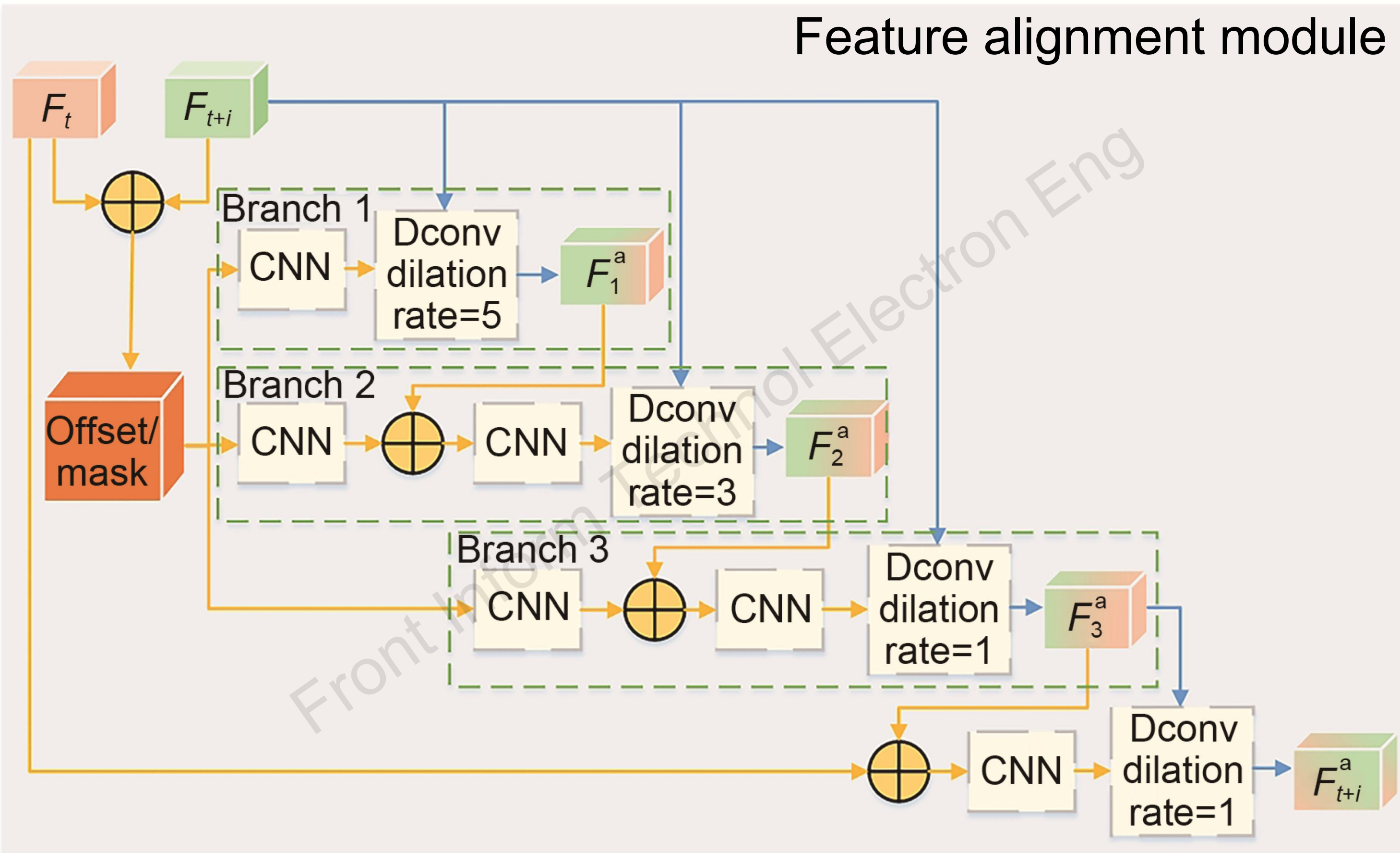
Overview of FaSRnet: at the feature level, auxiliary features are aligned with the current features and then fused with them through an attention mechanism; at the semantics level, the current heatmaps are refined using the difference information between the heatmaps as auxiliary features

Method

We propose a feature and semantics refinement network (FaSRnet) for human pose estimation, which uses temporal information to refine the output at the semantics and feature levels, and is intuitive. We design two components:

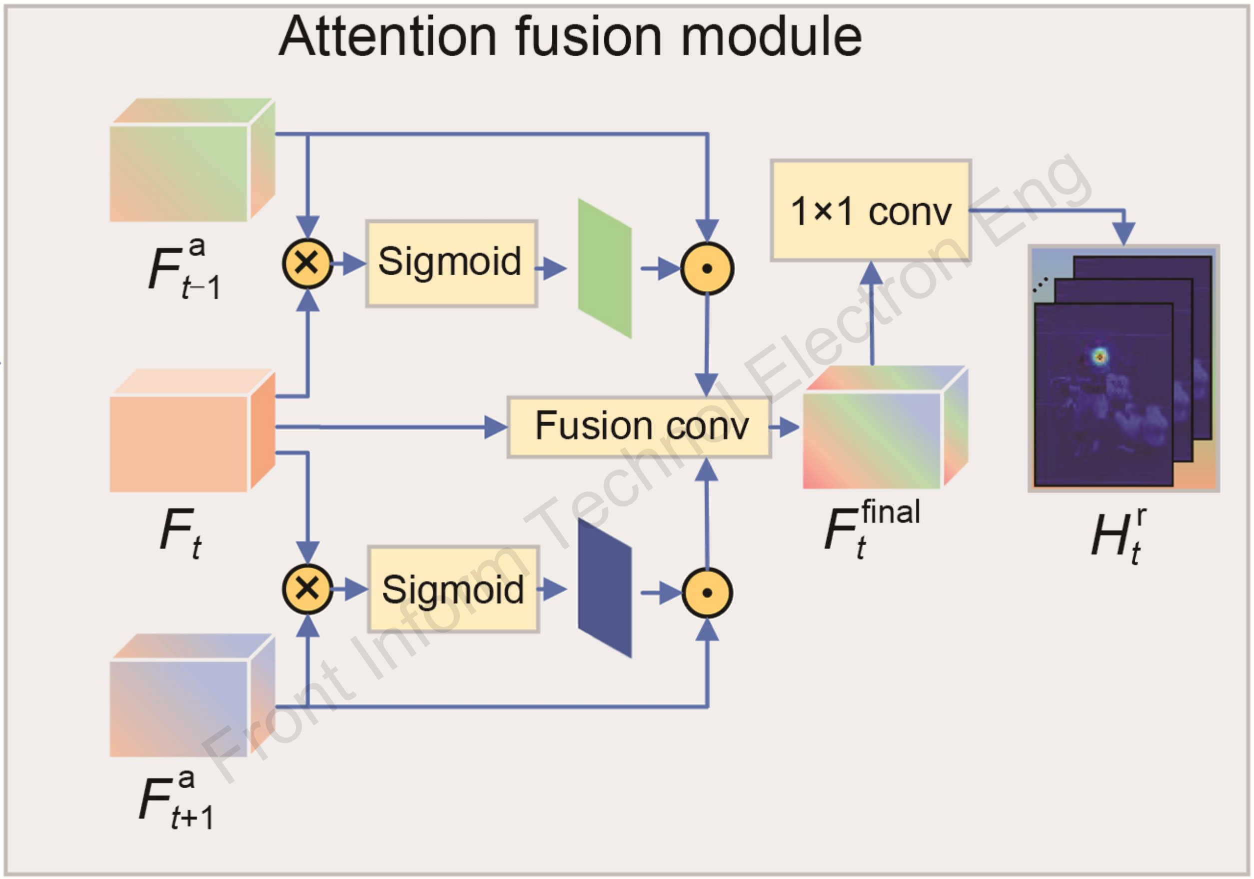
- feature refinement module (FRM), which is used to effectively aggregate key point information in adjacent features;
- semantics refinement module (SRM), which uses difference information from adjacent heatmaps to refine the current heatmaps.

Method



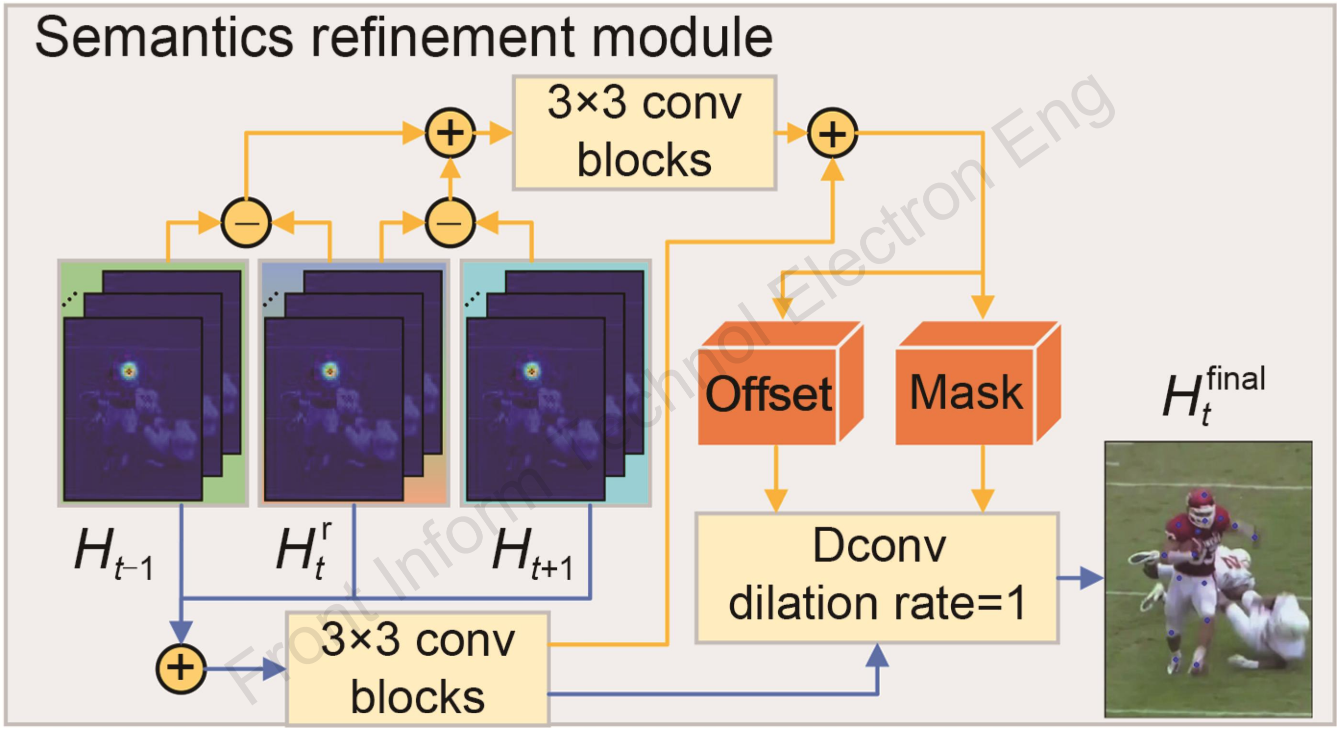
During feature alignment, this module repeatedly aligns auxiliary features with key features in different receptive fields to reduce the difference caused by feature distribution (the blue path).

Method



There is no guarantee that auxiliary features can help improve the current features. If wrong auxiliary features are fused, the quality of the current features and output accuracy will be reduced. Therefore, it is necessary to judge whether auxiliary features can help improve the refinement.

Method



The adjacent rough heatmaps contain obvious human body key point position information, the difference between adjacent heatmaps reflects the amount of human motion within two frames, and the difference information can effectively supplement the key point positioning error in the current frame due to motion blur and other reasons.

Method

The loss function is defined as

$$L_1 = \frac{1}{J} \sum_{j=1}^{J} V_j l_2 (H_j^{\mathrm{F}}, \hat{H}_j), \tag{17}$$

$$L_2 = \frac{1}{J} \sum_{j=1}^{J} V_j l_2 (H_j^{\mathrm{S}}, \hat{H}_j), \tag{18}$$

$$\mathrm{Loss} = \alpha L_1 + (1 - \alpha) L_2, \tag{19}$$

where H_j^{F} represents the heatmaps generated by 1×1 convolution after feature-level refinement, H_j^{S} represents the heatmaps generated after semantics-level refinement, $\hat{H}_j$ represents the ground-truth heatmaps, j is the key point number, V_j visualizes the key points in the label, and α is the weight coefficient of L_1, set to 0.4 in this method.

Major results

Table 1 Quantitative results of our method and state-of-the-art methods on the PoseTrack2017 validation set

Method	Year	AP							mAP
		Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	
PoseTracker	2018	67.5	70.2	62.0	51.7	60.7	58.7	49.8	60.6
PoseFlow	2018	66.7	73.3	68.3	61.1	67.5	67.0	61.3	66.5
JointFlow	2018	–	–	–	–	–	–	–	69.3
SimpleBaseline	2018	81.7	83.4	80.0	72.4	75.3	74.8	67.1	76.7
TML++	2019	–	–	–	–	–	–	–	71.5
FastPose	2019	80.0	80.3	69.5	59.1	71.4	67.5	59.4	70.3
STEmbedding	2019	83.8	81.6	77.1	70.0	77.4	74.5	70.8	77.0
HRNet	2019	82.1	83.6	80.4	73.3	75.5	75.3	68.5	77.3
MDPN	2019	85.2	88.5	83.9	77.5	79.0	77.0	71.4	80.7
PoseWarper	2019	81.4	88.3	83.9	78.0	82.4	80.5	73.6	81.2
Dynamic-GNN	2021	**88.4**	88.4	82.0	74.5	79.1	78.3	73.1	81.1
DCpose	2021	88.0	88.7	84.1	**78.4**	83.0	**81.4**	**74.2**	82.8
AlphaPose	2023	–	–	–	–	–	–	–	76.9
Ours	2022	88.1	**88.8**	**84.2**	**78.4**	**83.1**	**81.4**	**74.2**	**83.0**

The bold font denotes the best result

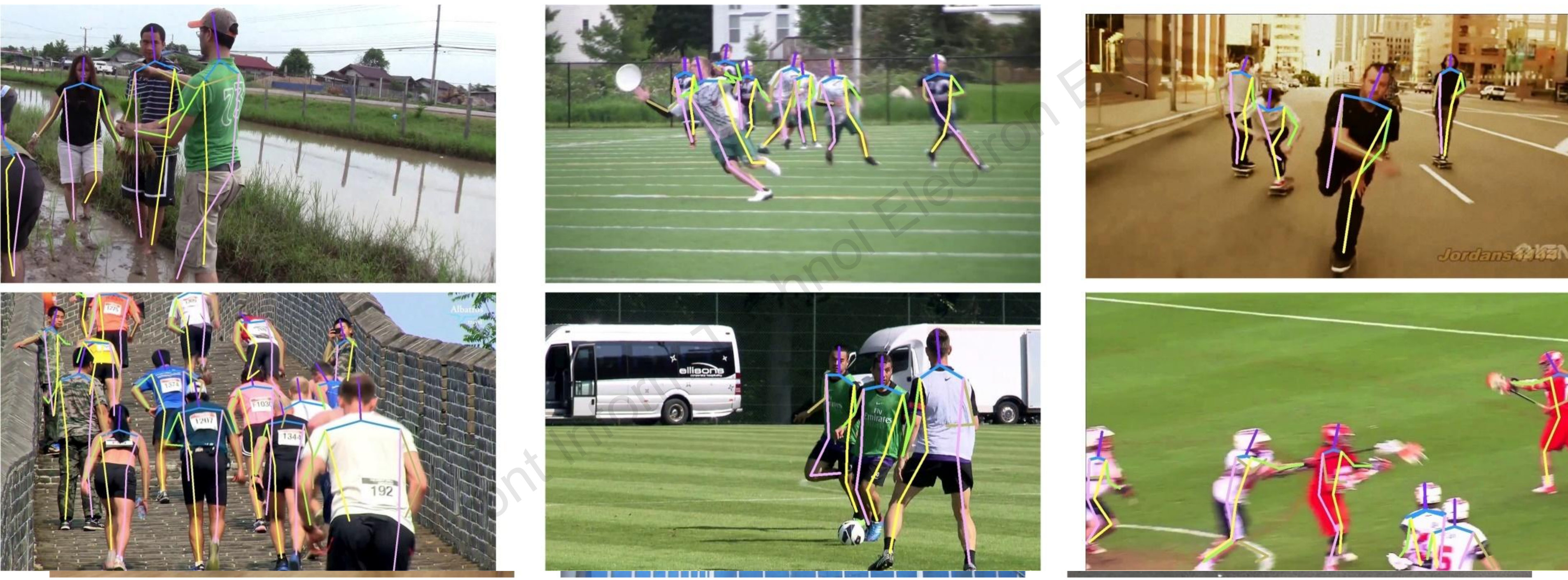
Major results

Table 2 Quantitative results of our method and state-of-the-art methods on the PoseTrack2017 test set

Method	Year	AP							mAP
		Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	
PoseTracker	2018	–	–	–	51.5	–	–	50.2	59.6
PoseFlow	2018	64.9	67.5	65.0	59.0	62.5	62.8	57.9	63.0
JointFlow	2018	–	–	–	53.1	–	–	50.4	63.4
SimpleBaseline	2018	80.1	80.2	76.9	71.5	72.5	72.4	65.7	74.6
TML++	2019	–	–	–	60.9	–	–	–	67.8
HRNet	2019	80.1	80.2	76.9	72.0	73.4	72.5	67.0	74.9
PoseWarper	2019	79.5	84.3	80.1	75.8	77.6	76.8	70.8	77.9
KeyTrack	2020	–	–	–	71.9	–	–	65.0	74.0
DCpose	2021	84.3	**84.9**	80.5	76.1	77.9	77.1	**71.2**	79.2
Ours	2022	**84.6**	84.8	**80.6**	**76.2**	**78.0**	**77.2**	71.0	**79.3**

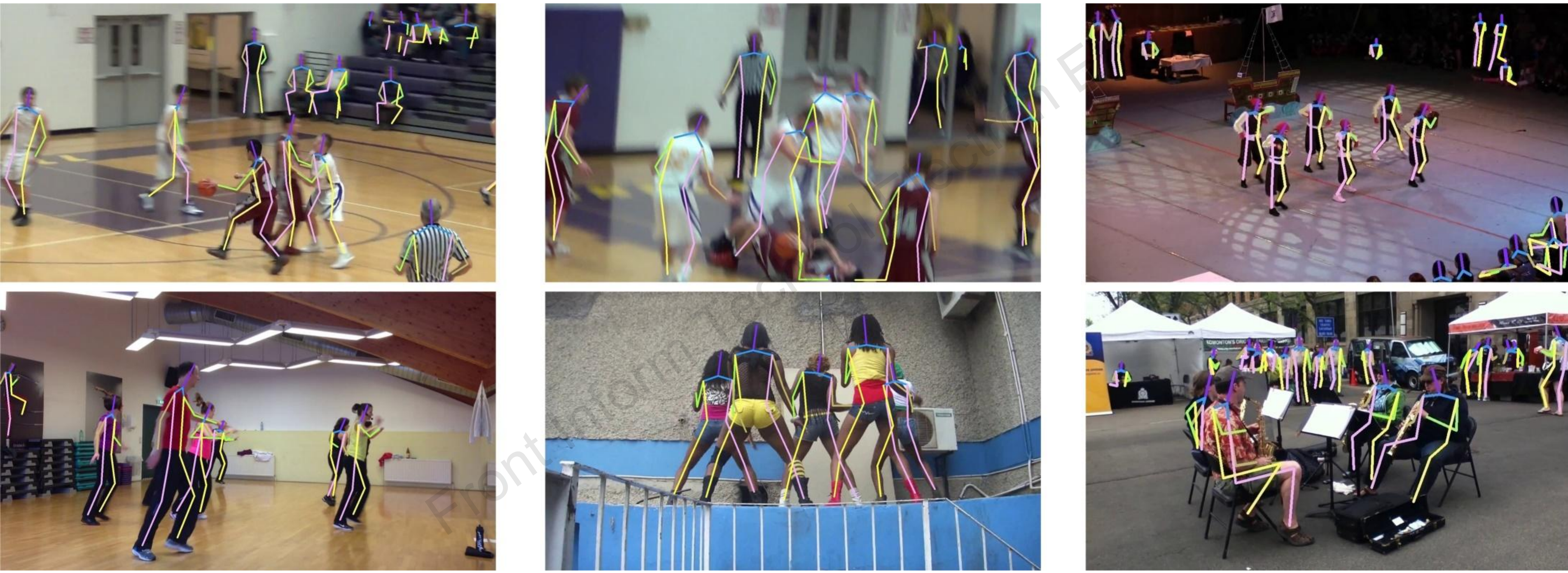
The bold font denotes the best result

Major results



Visualization results of some challenging scenarios in the PoseTrack2017 dataset. Scenes include motion blur, occlusion, and multiple persons

Major results



Visualization results of some challenging scenarios in the PoseTrack2018 dataset. Scenes include motion blur, occlusion, and multiple persons

Conclusions

We propose a video-based human pose estimation model. The method refines the current frame at feature and semantics levels. A multi-receptive field feature refinement module is designed to refine the predicted pose. Our semantics correction module uses the difference information between heatmaps to further refine the predicted pose. Our method has been validated on large-scale benchmark datasets PoseTrack2017 and PoseTrack2018, outperforming most existing methods.

Yuanhong ZHONG (Senior Member, IEEE) received the B.S. degree in communications engineering and the M.S. and Ph.D. degrees in communication and information systems from Chongqing University, Chongqing, China, in 2003, 2006, and 2011, respectively. He is currently an associate professor with School of Microelectronics and Communication Engineering, Chongqing University. His research interests include computer vision, machine learning, and intelligent unmanned systems.

Qianfeng XU received the B.S. degree in electronic information engineering and the M.S. degree in electronic science and technology from Chongqing University, Chongqing, China, in 2021 and 2023 respectively. His research interest is human pose estimation.

Daidi ZHONG received the Ph.D. degree from the Technical University of Tampere, Finland, in 2008. He is currently a professor at the School of Bioengineering, Chongqing University, Chongqing, China. He serves as the chairman of ISO/IEEE 11073-PHD, an expert in the field of WHO digital health, an academic member of ITU, director of the IEEE Standards Association, and a Chinese expert on ISO TC215. His current research interests include digital health and proactive health.

Xun YANG received the Ph.D. degree from Hefei University of Technology, Hefei, China, in 2017. From 2015 to 2017, he visited University of Technology Sydney (UTS), Australia, as a joint Ph.D. student. He was a research fellow with the NExT++ Research Center, National University of Singapore (NUS), from 2018 to 2021. He is currently a tenure-track professor with the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC). He regularly serves as a PC member and an invited reviewer for top-tier conferences and prestigious journals in multimedia and artificial intelligence, like IJCAI, AAAI, CVPR, ICCV, and *ACM Multimedia*. He also serves as an SPC member for AAAI 2021. His current research interests include information retrieval, cross-media analysis and reasoning, and computer vision.

Shanshan WANG received the B.E. and Ph.D. degrees from Chongqing University, Chongqing, China, in 2010 and 2020, respectively. She is currently an associate professor with the Institutes of Physical Science and Information Technology, Anhui University, Hefei, China. Her research interests include computer vision, domain adaptation, and data mining.