Supplementary Material: Temporal-Interim Pose Synthesis and Distillation for Dynamic Human Pose Estimation

Renjie Zhang, Di Lin, *Member, IEEE*, Xin Wang, Ruonan Liu, *Senior Member, IEEE*, Bin Sheng, *Member, IEEE*, George Baciu, *Senior Member, IEEE*, C. L. Philip Chen, *Life Fellow, IEEE*, and Ping Li, *Member, IEEE*

***Abstract*—This supplementary document provides additional details of our method, including ablation experimental results and extra visual results together with the ground truth images.**

I. Hyper Parameter Balancing Experiments

To illustrate the effectiveness of proposed loss terms. We conduct experiments with setting different values of hyper parameters of them. From the massive experiments, we observe that the values of γ_s will not influence directly to the internal of the $\mathcal{L}_{info}$. So we firstly conduct experiments based on different settings of γ_{info} and γ_s to find the best balancing of these two extra loss terms. From Table I, it is evident that without any information-level constraints ($\gamma_{info} = 0$), we rely solely on the localization error. This approach still yields better results than the original HRNet (77.3 mAP). From the first row, we can tell that the $\mathcal{L}_{info}$ is very important for the useful information extraction and the performance improvement. And the results from first column demonstrates that $\mathcal{L}_s$ can provide effective help for enhancing the motion information. From each column, we can tell that with the increasing of the value of γ_{info}, the mAP get higher and then down. The inappropriate hyperparameters can cause bad influence on the pose estimation. Considering all columns together, we will find that when $\gamma_{info} = 1$, the model will get the best results. The model performance changes related to γ_s have a similar trend with γ_{info}. And the best set value of the γ_s is 0.1.

TABLE I
Sensitivity to the proposed objectives. We report the results on PoseTrack2017 dataset.

mAP (γ_{info} \ γ_s)	0	0.1	1	5
0	83.0	83.5	83.9	83.7
0.1	85.4	85.9	85.3	85.2
1	85.6	**86.2**	85.8	85.5
5	85.5	85.8	85.7	85.3

In the Table II, we initialize the hyperparameters of the information theoretic objectives and empirically set the values of γ_{info} and γ_s as 1 and 0.1. By incorporating these objectives, we further enhance accuracy. The results from the first column ($\gamma_{com} = 0$), indicate that $\mathcal{L}_{var}$ and $\mathcal{L}_{red}$ may not be effective in the absence of $\mathcal{L}_{com}$. This is because $\mathcal{L}_{com}$ aims to maximize the task-relevant complementary information. But without $\mathcal{L}_{com}$, there is no additional useful information extracted. Under this condition, $\mathcal{L}_{red}$ will have no effect on the learning. The superior performance in the other columns validates the effectiveness of $\mathcal{L}_{com}$. The results from the first row of each block demonstrate that $\mathcal{L}_{var}$ aids in the specific part feature learning of our model, ensuring the feature diversity of each part group feature. From experiments of the last column, the last row of each block and the whole last block, the inappropriate hyperparameter of $\mathcal{L}_{com}$, $\mathcal{L}_{var}$ and $\mathcal{L}_{red}$ can also deteriorate the model performance, underscoring the importance of proper hyperparameter initialization.

II. Visual Results

To show the effectiveness of the pose synthesis, we provide more visual results of the synthesized poses in Fig. 1. Besides, we present more visual results of dynamic HPE on PoseTrack2017 [1] and PoseTrack2018 in Fig. 2.

References

[1] U. Iqbal, A. Milan, and J. Gall, "PoseTrack: Joint multi-person pose estimation and tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4654–4663.

Manuscript received 6 March 2024; revised 24 September 2024; accepted 6 February 2025. This work was supported in part by The Hong Kong Polytechnic University (PolyU) under Grant P0048387, Grant P0042740, Grant P0044520, Grant P0043906, Grant P0049586, and Grant P0050657, in part by the PolyU Research Institute for Sports Science and Technology under Grant P0044571, and in part by the National Natural Science Foundation of China under Grant 6247072353. *(Renjie Zhang and Di Lin contributed equally to this work.) (Corresponding author: Ping Li.)*

Renjie Zhang, Xin Wang, and George Baciu are with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong (e-mail: renjie.zhang@connect.polyu.hk; xin1025.wang@connect.polyu.hk; csgeorge@polyu.edu.hk).

Di Lin is with the College of Intelligence and Computing, Tianjin University, Tianjin 300072, China (e-mail: ande.lin1988@gmail.com).

Ruonan Liu is with the Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: ruonan.liu@sjtu.edu.cn).

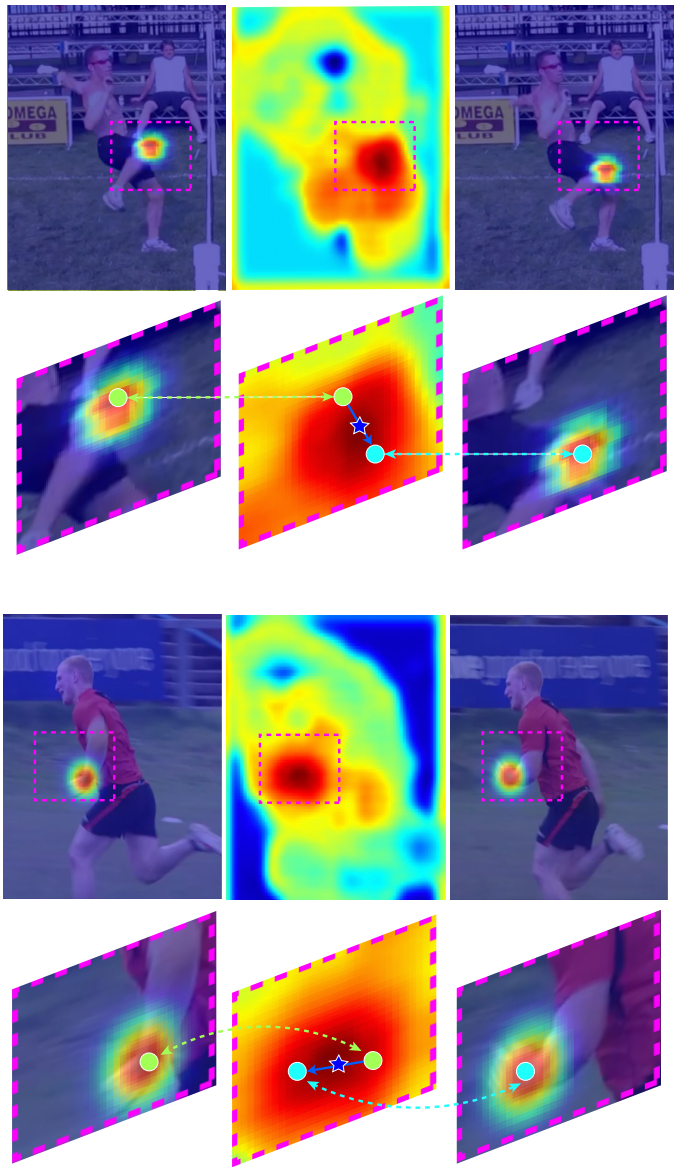
Bin Sheng is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: shengbin@sjtu.edu.cn).

C. L. Philip Chen is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China, and also with the Pazhou Lab, Guangzhou 510335, China (e-mail: philip.chen@ieee.org).

Ping Li is with the Department of Computing, the School of Design, and the Research Institute for Sports Science and Technology, The Hong Kong Polytechnic University, Hong Kong (e-mail: p.li@polyu.edu.hk).

TABLE II
SENSITIVITY TO THE PROPOSED INFORMATION-THEORETIC OBJECTIVES. WE REPORT THE RESULTS ON POSETRACK2017 DATASET.

mAP γ_{var} \ γ_{com}	0	0.1	1	5
$\gamma_{red} = 0$				
0	83.9	84.0	84.0	83.7
0.1	84.0	84.9	84.9	84.5
1	84.1	85.0	85.3	85.1
5	84.0	84.9	85.1	85.0
$\gamma_{red} = 0.1$				
0	83.9	85.5	85.0	84.8
0.1	84.0	85.4	85.4	85.3
1	83.9	85.6	85.9	85.6
5	84.0	85.3	85.5	85.5
$\gamma_{red} = 1$				
0	83.8	85.3	85.4	85.3
0.1	83.9	85.2	85.8	85.5
1	84.0	85.5	**86.2**	85.4
5	83.9	85.7	85.8	85.5
$\gamma_{red} = 5$				
0	83.8	84.9	85.1	85.4
0.1	83.8	85.0	85.2	85.6
1	83.9	85.2	85.4	85.7
5	84.0	85.3	85.2	85.6



(a) Previous Frame (b) Synthesized Heat Map (c) Next Frame

Fig. 1. Visual examples obtained by the pose synthesizer. The pink boxes select the areas with dynamic part locations. In the enlarged boxes in the second and fourth rows, we zoom in the parts. The green and cyan circles represent the locations of the same part in the adjacent frames. The blue arrows in the enlarged boxes represent the line trajectory formed by the parts in the previous frame and the next frame. The blue stars are the synthesized locations of parts in the short-time interval.

Fig. 2. The part recognition results of our method on benchmark datasets. We illustrate the challenging cases such as fast motion with blur, crowded background and kinds of pose occlusions.