Video-Based Human Pose Regression via Decoupled Space-Time Aggregation

Supplementary Material

Appendix

In the supplementary material, we provide:

A. Additional Implementation Details

Extracing Joint Embedding on HRNet. In the module of Joint-centric Feature Decoder (JFD), the feature embedding is extracted for each joint from the given global feature maps $\mathcal{F}_i(t')$ with $t' \in [t-T, t+T]$. In our implementation with HRNet backbones, specifically the HRNet-W48 variant, the high-resolution branch of the HRNet backbone is succeeded by a 1×1 convolutional layer (CONV) and a joint-wise fully connected feed-forward network (FFN) that encompasses a fully connected layer. The CONV layer consists of n kernels, where n represents the number of pose joints, resulting in a dedicated feature map for each joint. Subsequently, each individual feature map corresponding to a specific joint is flattened and fed into the shared FFN network, ultimately yielding its 32-bit feature embedding.

Dataset. We evaluate our models on three widely-utilized video-based benchmarks for human pose estimation: PoseTrack2017 [16], PoseTrack2018 [1], and PoseTrack21 [7]. Specifically, PoseTrack2017 includes 250 video clips for training and 50 videos for validation , with a total of 80, 144 pose annotations. PoseTrack2018 considerably increases the number of clips, containing 593 videos for training, 170 videos for validation, and a total of 153, 615 pose annotations. Both datasets identify 15 keypoints, with an additional label for joint visibility. The training videos are densely annotated in the center 30 frames, and validation videos are additionally labeled every four frames. PoseTrack21 further enriches and refines PoseTrack2018 especially for annotations of small persons and persons in crowds, including 177, 164 human pose annotations.

Optimization. We incorporate data augmentation including random rotation [-45°, 45°], random scale [0.65, 1.35], truncation (half body), and flipping during training. We adopt the AdamW optimizer [30] to train the entire network for 40 epochs, with a base learning rate of 2*e*-4, which is reduced by an order of magnitude at the 20*th* and 30*th* epochs. β_1 and β_2 are set to 0.9 and 0.999, respectively, and weight decay is set to 0.01.

Method	#Params	GFLOPs of Backbone	GFLOPs of Net. Head	mAP
heatmap-based				
PoseWarper [2]	39.1M	4.1	90.3	75.9
DCPose [22]	35.6M	4.1	21.7	77.1
regression-based				
DSTA (Ours)	**24.6M**	4.1	**0.01**	**78.6**

Table 8. **Computation complexity** with ResNet-50 backbone. #Params includes the parameters of entire network. All methods utilize the same two auxiliary frames as in [22].

Method	#Params	GFLOPs of Backbone	GFLOPs of Net. Head	mAP
heatmap-based				
PoseWarper [2]	14.8M	0.35	73.5	67.7
DCPose [22]	11.3M	0.35	4.9	68.8
regression-based				
DSTA (Ours)	**2.4M**	0.35	**0.01**	**71.0**

Table 9. **Computation complexity** with MobileNet-V2 backbone. #Params includes the parameters of entire network. All methods utilize the same two auxiliary frames as in [22].

B. Computation Complexity on More Backbones

Tables 8 and 9 present additional comparisons of computation complexity between our regression-based method and the heatmap-based methods. These experiments were conducted on the PoseTrack2017 validation set using the ResNet-50 and MobileNet-V2 backbones, respectively. While implementing the heatmap-based PoseWarper [2] and DCPose [22], we utilized their official open-source codes. In their network heads, similar to SimpleBase [37], we employed 3 deconvolution layers to generate high-resolution heatmaps from the backbones.

As shown, our method outperforms heatmap-based methods in both backbones, while utilizing significantly lower computation complexity and fewer model parameters. In addition, when compared to the HRNet backbone's results presented in Table 3 of the main paper, our method achieves even greater savings in computational costs and

Method	Bkbone	Head	Should.	Elbow	Wrist	Hip	Knee	Ankle	**Mean**
heatmap-based									
AlphaPose [8]	ResNet-50	63.9	78.7	77.4	71.0	73.7	73.0	69.7	71.9
MDPN [12]	ResNet-152	75.4	81.2	79.0	74.1	72.4	73.0	69.9	75.0
Dyn.-GNN [42]	HRNet-W48	80.6	84.5	80.6	74.4	75.0	76.7	71.8	77.9
PoseWarp. [2]	HRNet-W48	79.9	86.3	82.4	77.5	79.8	78.8	73.2	79.7
PT-CPN++ [44]	CPN [4]	82.4	**88.8**	**86.2**	**79.4**	72.0	80.6	**76.2**	80.9
DCPose [22]	HRNet-W48	84.0	86.6	82.7	78.0	80.4	79.3	73.8	80.9
DetTrack [35]	HRNet-W48	84.9	87.4	84.8	79.2	77.6	79.7	75.3	81.5
FAMIPose [23]	HRNet-W48	85.5	87.7	84.2	79.2	81.4	**81.1**	74.9	**82.2**
regression-based									
DSTA (Ours)	ResNet-152	85.2	87.1	80.5	74.4	79.6	78.0	69.7	79.6
DSTA (Ours)	HRNet-W48	**86.2**	**88.6**	84.2	78.5	**82.0**	79.2	73.7	**82.1**
DSTA (Ours)	ViT-H	85.9	88.8	85.0	81.1	81.5	83.0	77.4	83.4

Table 10. **Comparison with the SOTA** on PoseTrack2018 val. set. Similar to FAMI-Pose [23], our proposed DSTA sets the temporal span T to 2, consisting of two preceding and two subsequent frames, totalling four auxiliary frames.

Method	Bkbone	Head	Should.	Elbow	Wrist	Hip	Knee	Ankle	**Mean**
heatmap-based									
SimBase. [37]	ResNet-152	80.5	81.2	73.2	64.8	73.9	72.7	67.7	73.9
HRNet [28]	HRNet-W48	81.5	83.2	81.1	75.4	79.2	77.8	71.9	78.8
PoseWarp. [2]	HRNet-W48	82.3	84.0	82.2	75.5	80.7	78.7	71.6	79.5
DCPose [22]	HRNet-W48	83.7	84.4	82.6	**78.7**	80.1	79.8	74.4	80.7
FAMIPose [23]	HRNet-W48	83.3	85.4	82.9	78.6	81.3	**80.5**	**75.3**	81.2
regression-based									
DSTA (Ours)	ResNet-152	86.1	85.5	80.0	74.6	80.5	76.9	70.2	79.6
DSTA (Ours)	HRNet-W48	**87.5**	**86.6**	**83.3**	**78.7**	**82.7**	78.3	73.9	**82.0**
DSTA (Ours)	ViT-H	87.5	87.0	84.2	81.4	82.3	82.5	77.7	83.5

Table 11. **Comparison with the SOTA** on PoseTrack21 val. set. Similar to FAMI-Pose [23], our proposed DSTA sets the temporal span T to 2, consisting of two preceding and two subsequent frames, totalling four auxiliary frames.

#Token Size	8	16	32	64
mAP	76.1	78.0	78.6	78.6

Table 12. **Different sizes of joint tokens**. In the experimental setup, we utilized the ResNet-50 backbone along with two auxiliary frames.

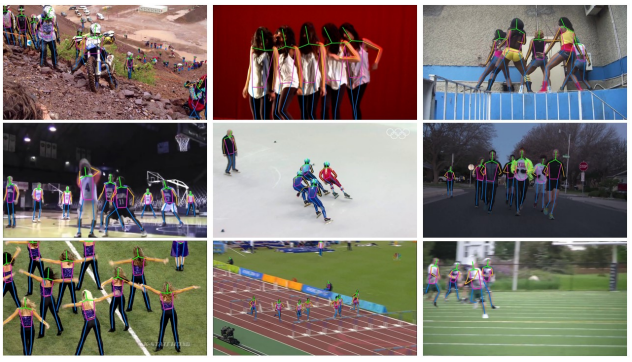

Figure 4. **Additional qualitative results** of our DSTA on the PoseTrack datasets.

model parameters on these smaller backbone networks. For instance, when utilizing the MobileNet-V2 backbone, our regression-based network incorporates a mere **2.4** million parameters, whereas the heatmap-based networks demand a significantly higher number, specifically 14.8 million and 11.3 million parameters. On the other hand, when employing the ResNet-50 backbone, the FLOPs of our regression-based head are almost negligible, accounting for just **1/9030** or **1/2170** of those required by the heatmap-based heads. The superior computational and storage efficiency of our proposed regression framework holds immense value in the industry, especially for edge devices and real-time video applications.

C. Experiments on PoseTrack2018/21 Datasets

Tables 10 and 11 present the comparisons of our method with the state-of-the-art methods on the PoseTrack2018 and PoseTrack21 valdation sets, respectively. These results further demonstrate that our proposed regression-based method achieves performance that is either superior to, or at the very least, on par with the state-of-the-art heatmap-based methods.

D. Additional Ablation Study

Size of Joint Tokens. In this additional study, we conduct experiemnts to examine the influence of the adopted size of the joints' feature embedding (*i.e.*, joint token). Table 12 presents the performance variations resulting from different joint token sizes on the PoseTrack2017 validation set. As the size of the joint tokens increases, a gradual improvement in performance can be observed. However, beyond a size of 16, the performance tends to plateau, suggesting that further increases in token size do not yield commensurate improvements. This indicates that each pose joint requires a sufficiently large feature token to store its relevant feature information, but a too large feature token will only cause spatial redundancy. Therefore, in our experiments, we have opted to use a token size of 32, striking a balance between capturing sufficient feature information and avoiding unnecessary spatial redundancy.

E. Qualitative Results

Additional qualitative results on PoseTrack datasets are shown in Fig. 4. Additional results can be found in the accompanying video material.