

Beyond Static Frames: Temporal Aggregate-and-Restore Vision Transformer for Human Pose Estimation

Supplementary Material

Appendix

In the supplementary material, we provide:

- §A Additional Implementation Details.
- §B Experiments on PoseTrack2018/21 Datasets.
- §C Additional Ablation Study.
- §D Qualitative Results.
- §E Limitation and Future Work.

A. Additional Implementation Details

Dataset. Our models are evaluated on three widely used video-based human pose estimation benchmarks: PoseTrack2017 [16], PoseTrack2018 [1], and PoseTrack21 [7]. Together, these datasets form a comprehensive evaluation suite that spans diverse scenarios, motion patterns, and levels of visual complexity. A brief overview of each dataset is provided below.

- **PoseTrack2017** [16] contains 250 training videos and 50 validation videos, with a total of 80,144 pose annotations. Each person instance is annotated with 15 keypoints and corresponding visibility labels. Training clips are densely annotated in the central 30 frames, whereas validation clips are annotated every four frames.
- **PoseTrack2018** [1] substantially enlarges the dataset, providing 593 training videos and 170 validation videos, amounting to 153,615 pose annotations. It uses the same 15-keypoint annotation scheme and visibility labels as PoseTrack2017, with an identical annotation protocol—dense annotations in the central 30 frames of training videos and sparser annotations (every four frames) for validation videos.
- **PoseTrack21** [7] extends and refines PoseTrack2018, particularly improving annotations for small individuals and people in crowded scenes. It includes 177,164 human pose annotations while preserving the same keypoint definition and annotation strategy. The updates make it more challenging, especially regarding occlusion and scale variations.

Optimization. We implemented TAR-ViTPose in PyTorch. Data augmentation includes random scaling within the range $[0.65, 1.35]$, random rotation within $[-45^\circ, 45^\circ]$, random cropping, and horizontal flipping. The input image resolution is set to 384×288 . The initial learning rate is 5×10^{-6} and is reduced by 50% every 5 epochs. We use the AdamW optimizer for model training.

Method	Head	Should.	Elbow	Wrist	Hip	Knee	Ankle	Mean
<i>Backbone: ViT-S</i>								
ViTPose [41]	80.6	84.1	79.9	73.5	77.0	76.2	70.3	78.2
TAR-ViTPose(Ours)	82.8	85.6	81.5	75.2	78.0	78.1	72.1	79.3
<i>Backbone: ViT-B</i>								
ViTPose [41]	83.2	86.3	82.8	77.5	78.2	79.4	74.9	80.5
TAR-ViTPose(Ours)	84.3	87.8	83.6	79.4	79.3	82.8	76.5	82.1
<i>Backbone: ViT-H</i>								
ViTPose [41]	85.4	87.0	84.7	81.0	79.1	82.0	76.5	82.4
TAR-ViTPose(Ours)	86.8	88.9	86.1	81.9	82.9	83.4	77.9	84.2

Table 7. Comparison with the ViTPose baseline (mAP) on PoseTrack2018 val. set.

Method	Head	Should.	Elbow	Wrist	Hip	Knee	Ankle	Mean
<i>Backbone: ViT-S</i>								
ViTPose [41]	79.7	84.4	80.4	73.8	77.2	77.3	71.2	77.8
TAR-ViTPose(Ours)	81.6	84.7	80.9	74.2	77.8	77.9	71.7	78.5
<i>Backbone: ViT-B</i>								
ViTPose [41]	81.8	85.2	82.1	76.6	78.5	79.5	74.5	79.9
TAR-ViTPose(Ours)	82.7	85.9	83.5	79.2	79.5	81.7	76.8	81.4
<i>Backbone: ViT-H</i>								
ViTPose [41]	83.2	87.0	84.1	80.4	79.5	81.9	77.3	82.0
TAR-ViTPose(Ours)	86.6	88.3	85.3	82.0	83.7	83.4	78.2	84.1

Table 8. Comparison with the ViTPose baseline (mAP) on PoseTrack21 val. set.

B. Experiments on PoseTrack2018/21 Datasets

B1. Comparison with the ViTPose Baseline

Tables 7 and 8 present comparisons between our method and the state-of-the-art ViT-based single-frame baseline ViTPose [41] on the PoseTrack2018 and PoseTrack21 validation sets, respectively. The results further demonstrate that our video-based approach consistently achieves substantial performance gains across all ViT backbones, highlighting the importance of exploiting temporal cues from adjacent frames, which single-frame baselines such as ViTPose are inherently unable to capture.

B2. Comparison with State-of-the-Art Methods

As shown in Tables 9 and 10, our approach establishes new state-of-the-art performance on the PoseTrack2018 and PoseTrack21 datasets, achieving **84.2** mAP and **84.1** mAP, respectively. When evaluated with ground-truth bounding boxes, our method attains 89.8 mAP and 91.0 mAP, yielding additional gains of **5.6** and **6.9** points. These results consistently outperform the state-of-the-art video-based methods MTPose [31] and Poseidon [27], both of which also rely on ground-truth bounding boxes.

Method	Backbone	Head	Should.	Elbow	Wrist	Hip	Knee	Ankle	Mean
<i>Using bounding boxes predicted by the Faster R-CNN detector [28]</i>									
PoseWarper [2]	HRNet-W48	79.9	86.3	82.4	77.5	79.8	78.8	73.2	79.7
PAVE-Net† [44]	Swin-L	84.9	87.5	81.7	74.5	77.9	76.9	72.4	79.7
DCPose [24]	HRNet-W48	84.0	86.6	82.7	78.0	80.4	79.3	73.8	80.9
FAMI-Pose [25]	HRNet-W48	85.5	87.7	84.2	79.2	81.4	81.1	74.9	82.2
DSTA [14]	ViT-H	85.9	88.8	85.0	81.1	81.5	83.0	77.4	83.4
TAR-ViTpose	ViT-S	82.8	85.6	81.5	75.2	78.0	78.1	72.1	79.3
TAR-ViTpose	ViT-B	84.3	87.8	83.6	79.4	79.3	82.8	76.5	82.1
TAR-ViTpose	ViT-H	86.8	88.9	86.1	81.9	82.9	83.4	77.9	84.2
<i>Using bounding boxes from unspecified sources or ground-truth ones (*)</i>									
Dyn.-GNN [43]	HRNet-W48	80.6	84.5	80.6	74.4	75.0	76.7	71.8	77.9
DetTrack [37]	HRNet-W48	84.9	87.4	84.8	79.2	77.6	79.7	75.3	81.5
TDMI [10]	HRNet-W48	86.2	88.7	85.4	80.6	82.4	82.1	77.5	83.5
DiffPose [9]	ViT (#)	85.0	87.7	84.3	81.5	81.4	82.9	77.6	83.0
MTPose* [31]	ViT (#)	89.4	92.4	90.1	87.3	85.7	89.7	88.1	89.0
CM-Pose [4]	ViT (#)	85.7	88.9	85.8	81.0	84.4	84.2	80.1	84.4
GLSMamba [11]	ViT-H	85.6	88.9	86.5	83.6	82.9	85.7	81.4	84.9
Poseidon* [27]	ViT-H	88.8	91.4	88.6	86.3	83.3	88.8	87.2	87.8
TAR-ViTpose*	ViT-S	85.6	88.5	84.2	79.2	80.8	81.9	78.7	82.9
TAR-ViTpose*	ViT-B	88.7	91.8	89.4	86.1	83.6	88.5	87.3	88.0
TAR-ViTpose*	ViT-H	90.9	93.1	91.1	88.5	84.7	90.4	89.0	89.8

Table 9. Comparison with the SOTAs on PoseTrack2018 val. set. ‘#’ indicates that the ViT backbone used is not specified, ‘†’ indicates the end-to-end approach. Similar to FAMI-Pose [25] and DSTA [14], our proposed TAR-ViTpose sets the temporal span T to 2, which includes two preceding and two succeeding frames, totaling four auxiliary frames.

C. Additional Ablation Study

Mask Threshold. We further analyze the influence of the mask threshold ϕ in Eq. 4 on constructing the joint-specific binary masks. As shown in Table 11, masking thresholds in the range of 0.1 to 0.25 all yield high human pose estimation accuracy. Intuitively, a lower threshold enlarges the attention region around each joint’s location in successive frames, while a higher threshold produces a smaller and more concentrated focus area. With a threshold of 0.2, the model achieves an optimal balance, reaching 84.0 mAP.

Number of Layers in JTA. Table 12 analyzes the effect of varying the number of layers in our JTA(\cdot, \cdot) module. Increasing the depth from 2 to 6 layers consistently improves performance (83.4 \rightarrow 84.0 mAP), which is consistent with the general observation that deeper architectures offer stronger representation capacity. However, further increasing the number of layers beyond 6 yields no additional benefit and even leads to a slight performance drop (83.9 mAP with 8 or 10 layers). We attribute this saturation to the redundancy introduced by excessive temporal aggregation, which may over-smooth or dilute frame-specific cues. Overall, a depth of 6 layers provides the best balance between accuracy and efficiency, and is therefore adopted as the default configuration in our main experiments.

Method	Backbone	Head	Should.	Elbow	Wrist	Hip	Knee	Ankle	Mean
<i>Using bounding boxes predicted by the Faster R-CNN detector [28]</i>									
SimBase. [39]	ResNet-152	80.5	81.2	73.2	64.8	73.9	72.7	67.7	73.9
HRNet [32]	HRNet-W48	81.5	83.2	81.1	75.4	79.2	77.8	71.9	78.8
PAVE-Net † [44]	Swin-L	84.7	86.5	81.9	74.7	77.4	76.6	71.4	79.4
PoseWarper [2]	HRNet-W48	82.3	84.0	82.2	75.5	80.7	78.7	71.6	79.5
DCPose [24]	HRNet-W48	83.7	84.4	82.6	78.7	80.1	79.8	74.4	80.7
FAMI-Pose [25]	HRNet-W48	83.3	85.4	82.9	78.6	81.3	80.5	75.3	81.2
DSTA [14]	ViT-H	87.5	87.0	84.2	81.4	82.3	82.5	77.7	83.5
TAR-ViTpose	ViT-S	79.0	83.1	79.8	73.6	75.3	77.2	72.6	77.4
TAR-ViTpose	ViT-B	82.7	85.9	83.5	79.2	79.5	81.7	76.8	81.4
TAR-ViTpose	ViT-H	86.6	88.3	85.3	82.0	83.7	83.4	78.2	84.1
<i>Using bounding boxes from unspecified sources or ground-truth ones (*)</i>									
TDMI [10]	HRNet-W48	85.8	87.5	85.1	81.2	83.5	82.4	77.9	83.5
DiffPose [9]	ViT (#)	84.7	85.6	83.6	80.8	81.4	83.5	80.0	82.9
MTPose* [31]	ViT (#)	92.0	91.7	88.7	85.5	86.4	86.6	85.3	88.3
CM-Pose [4]	ViT (#)	88.9	88.3	84.4	81.9	84.6	83.7	78.8	84.3
GLSMamba [11]	ViT-H	87.0	86.9	85.4	83.2	83.4	84.8	80.8	84.7
Poseidon* [27]	ViT-H	92.2	90.8	88.3	85.8	85.5	87.7	85.7	88.3
TAR-ViTpose*	ViT-S	81.6	84.7	80.9	74.2	77.8	77.9	71.7	78.5
TAR-ViTpose*	ViT-B	82.9	87.8	85.9	82.1	80.1	84.5	82.0	83.6
TAR-ViTpose*	ViT-H	94.3	93.3	91.3	88.8	88.7	90.2	88.6	91.0

Table 10. Comparison with the SOTAs on PoseTrack21 val. set. ‘#’ indicates that the ViT backbone used is not specified, ‘†’ indicates the end-to-end approach. Similar to FAMI-Pose [25] and DSTA [14], our proposed TAR-ViTpose sets the temporal span T to 2, which includes two preceding and two succeeding frames, totaling four auxiliary frames.

Threshold ϕ	0.1	0.15	0.2	0.25	0.3
mAP	83.2	83.7	84.0	83.5	82.9

Table 11. Impact of different mask thresholds.

#Layer	2	4	6	8	10
mAP	83.4	83.8	84.0	83.9	83.9

Table 12. Different number of layers in JTA.

#Layer	1	2	3	4
mAP	84.0	83.9	84.0	84.0

Table 13. Different number of layers in GRA.

Number of Layers in GRA. In our GRA module, we use a single cross-attention layer to inject the aggregated temporal information back into the feature space of the current frame. As shown in Table 13, using only one layer already achieves strong performance (84.0 mAP), and increasing the number of layers does not provide any additional improvement. We attribute this to the nature of GRA: it does not extract or aggregate new features but simply restores the joint-specific temporal features to the current-frame representation, a task that can be effectively accomplished with a shallow design. Therefore, we adopt a single cross-attention layer as the default configuration to maintain high accuracy.



Figure 5. Additional qualitative results of our TAR-ViT-Pose. The first two rows are from the PoseTrack datasets, while the last row is from in-the-wild videos. Across all examples, the model remains robust under occlusion, motion blur, complex poses, and defocus.

#Auxiliary Frame	ViT-S	ViT-B	ViT-H
1 $\{-1\}$	81.4	83.2	86.1
2 $\{-1, +1\}$	81.5	83.4	86.3
4 $\{-2, -1, +1, +2\}$	81.9	84.0	86.8

Table 14. Different number of auxiliary frames. ‘-’ indicates previous frames while ‘+’ indicates subsequent frames.

while achieving optimal efficiency.

Auxiliary Frame. In addition, we investigate the impact of using different numbers of auxiliary frames. The results reported in Table 14 consistently show that increasing the number of auxiliary frames leads to improved performance across ViT backbones of different scales. This observation aligns with our intuition that incorporating more auxiliary frames provides richer and more complementary temporal information, thereby enabling more accurate and robust pose estimation for the key frame.

D. Qualitative Results

D1. Additional qualitative results on the PoseTrack validation sets and in-the-wild videos are presented in Fig. 5. More results are provided in the accompanying video.

D2. To further validate that the proposed mask-aware atten-

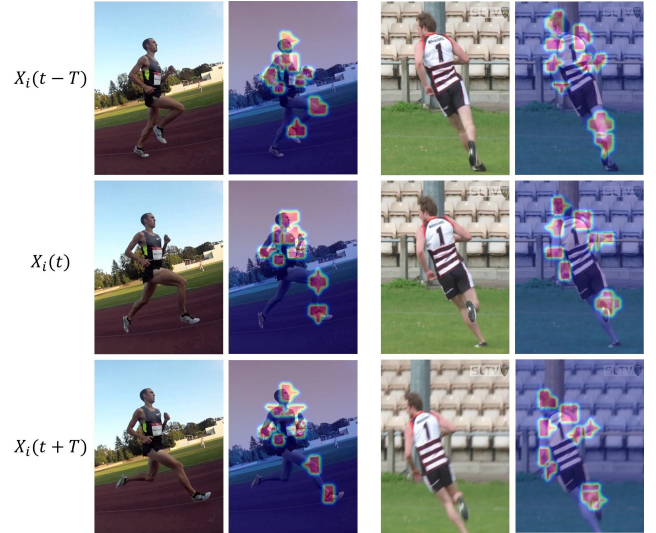


Figure 6. Visualization of attention heatmaps for joint query tokens with mask-aware attention. For each example, given a current frame $X_i(t)$ and its neighboring frames $X_i(t-T)$ and $X_i(t+T)$, we visualize the attention heatmaps of nine joint query tokens, including the head, shoulders, elbows, wrists, right knee, and right ankle, over the feature maps of these three frames.

tion enables each joint query token to effectively focus on its corresponding keypoint regions across frames while sup-

pressing interference from unrelated areas, additional attention heatmaps of joint query tokens are provided in Fig. 6.

E. Limitation and Future Work

It is important to clarify that our work does not target temporal pose tracking. Instead, it introduces a simple yet robust temporal Vision Transformer framework that delivers strong performance for 2D pose estimation in videos. Because the method does not explicitly impose temporal consistency across frames, it may occasionally produce slight temporal inconsistencies, especially in heavily occluded scenes, as shown in the supplementary video. Looking ahead, we plan to extend our framework to multi-person pose tracking by incorporating temporal identity consistency.