NCSTR: Node-Centric Decoupled Spatio-Temporal Reasoning for Video-based Human Pose Estimation

Quang Dang Huynh Xuefei Yin Andrew Busch Hugo G. Espinosa
Alan Wee-Chung Liew Matthew T.O. Worsey Yanming Zhu
Griffith University, Gold Coast, QLD, Australia
quangdang.huynh@griffith.edu.au

NCSTR: Node-Centric Decoupled Spatio-Temporal Reasoning for Video-based Human Pose Estimation

Supplementary Material

1. Key Implementation Details

We adopt a Vision Transformer based single frame pose estimator as our backbone network, following ViTPose [4]. The backbone is initialized from publicly available COCO pretrained weights via the MMPose framework [2]. For the proposed model, we train for 15 epochs using the AdamW optimizer and a multi-step learning rate schedule with an initial learning rate of 2×10^{-4}, decayed by a factor of 0.1 at epochs 5 and 11. Our NCSTR is implemented in PyTorch, and all training and evaluation are conducted on an NVIDIA A100-SXM4-80GB GPU.

2. Efficiency and Runtime Analysis

All timings use strictly causal inference (T=3: two past frames plus current frame), input resolution 384×288, on a single NVIDIA RTX 3090 GPU, with inference batch size B=1 (one pose-network forward pass per current-frame output). Tab. 1 summarizes the accuracy–efficiency trade-off across backbone variants. Smaller backbones reduce latency and peak memory substantially while remaining competitive, whereas ViTPose-Huge maximizes accuracy at higher compute cost.

Table 1. Efficiency breakdown across backbone variants. FPS and peak memory measured at batch size 1, causal inference (T=3, 384×288), single RTX 3090 GPU.

Backbone	mAP ↑	FPS ↑	Latency (ms) ↓	Peak Mem. (MB) ↓
ViTPose-H	86.4	19.2	52.1	2980
ViTPose-L	85.9	21.9	45.7	1530
ViTPose-B	83.8	23.8	42.0	722
ViTPose-S	82.3	25.3	39.5	462

3. Replacement Ablations

The component ablations in the main paper (Table 5) demonstrate the contribution of each module via removal studies. Here we provide complementary *replacement ablations*: capacity-preserving swaps that test whether specific design choices are essential or interchangeable with simpler standard alternatives. All experiments are conducted on PoseTrack17 using the ViTPose-Huge backbone.

Global 2-hop vs. fully-connected. The global spatial branch uses a 2-hop expanded skeleton graph rather than dense (fully-connected) attention. Replacing with full connectivity reduces mAP from 86.4 to 85.7 (-0.7), showing that "more connectivity" is not automatically better and that controlled 2-hop expansion is a better inductive bias than full density.

NSEF gating vs. sum fusion. The Node-Space Expert Fusion (NSEF) module combines local and global node predictions via a learned convex gate. Replacing it with direct summation drops mAP from 86.4 to 84.7 (-1.7), the largest single drop in this table, confirming that adaptive per-joint gating in node space is not interchangeable with a standard fusion baseline.

Temporal–spatial order: sequential vs. parallel. Our default design performs temporal propagation first, then spatial constraint reasoning (sequentially). Replacing this with parallel temporal and spatial branches followed by concatenation reduces mAP to 85.8 (-0.6), indicating that the benefit comes not simply from including both operations but from the specific sequential refinement schedule.

GAT vs. MLP. We choose GAT because its topology-constrained message passing matches the pose graph structure: one node per joint, edges from skeletal connectivity, and adaptive attention weights within a fixed neighbor set. This allows the model to emphasize informative joint relations under occlusion and fast motion while preserving anatomical locality, and it naturally supports our 1-hop local and 2-hop global designs. In contrast, MLPs do not model structured joint interactions. Replacing GAT with an MLP drops mAP from 86.4 to 84.0 (-2.4), providing the strongest single-component evidence for topology-controlled message passing.

Table 2. Replacement ablations on PoseTrack17. Each row replaces one design choice with a standard alternative while keeping the rest of the pipeline unchanged. Negative Δ indicates worse performance than the full model (86.4 mAP).

Method / Replacement	mAP ↑	Δ
Full model (default)	86.4	–
Global 2-hop → fully-connected	85.7	-0.7
NSEF gating → sum fusion	84.7	-1.7
Temporal then spatial → in parallel	85.8	-0.6
GAT reasoning → MLP (no graph)	84.0	-2.4

4. Effect of Bounding Box Detector on PoseTrack17

It is common in the PoseTrack literature for methods not to release their bounding box files or fully specify the detector

used, making strict comparisons across methods difficult. Following DCPose [1], we use the YOLOv3 [3] bounding box files provided by DCPose as a fixed, stable detection benchmark that has been adopted by multiple prior works. This ensures that performance differences reflect the pose model rather than detector-specific engineering.

To quantify how much of any remaining gap is attributable to bounding boxes rather than pose reasoning, we conducted diagnostic experiments using our method with different detection inputs:

Table 3. Effect of bounding box quality on PoseTrack17 mAP (ViTPose-Huge backbone). Detection quality can explain gaps between top-performing methods that do not disclose their detector.

Detection Input	mAP ↑
YOLOv3 boxes	86.4
YOLO26 boxes	87.5
Ground-truth boxes	89.5

The large increase under GT boxes confirms that detection quality can dominate small differences among top-performing methods on PoseTrack17, and that the pose model itself is not the limiting factor. The remaining gap versus TIPose [5] (86.7 mAP) falls within the range attributable to undisclosed detector protocols.

5. Local vs. Global Branch Differences

The local and global branches in DSTAG share the same high-level design (joint embedding → graph learning) but differ in three concrete technical ways, as summarized below.

Table 4. Structural differences between the local and global branches.

	PQE mask	Spatial adjacency	Functionality
Local branch	Local radius	1-hop skeleton	Fine-grained anatomical structure
Global branch	Global radius	2-hop skeleton	Long-range structural context

PQE mask. In the Pose-Query Encoder, the local branch attends within a tighter velocity-adaptive window ($r_{t,j}^{\text{local}} = 2r_{t,j} + 1$), while the global branch uses a slightly larger window ($r_{t,j}^{\text{global}} = r_{t,j}^{\text{local}} + 4$). Both are centered at the velocity-extrapolated joint position.

Spatial adjacency. The local spatial GAT uses the standard 1-hop skeletal adjacency $\mathcal{A}^{\text{local}}$ (each joint aggregates only from its direct anatomical neighbors), enforcing fine-grained anatomical locality. The global spatial GAT uses a 2-hop expanded skeleton $\mathcal{A}^{\text{global}}$ (each joint can additionally attend to neighbors-of-neighbors), providing wider structural context without resorting to fully dense attention. The replacement ablation (global 2-hop → fully-connected, -0.7 mAP; Tab. 2) confirms that the controlled expansion is a better design than unconstrained density.

Cross-branch attention. After independent spatial reasoning, a bi-directional cross-branch attention step allows the two branches to exchange information: global features attend to local features and vice versa. This reciprocal exchange ensures that each joint's final representation benefits from both fine-grained and long-range structural cues before expert fusion.

References

[1] Zhenguang Liu, Haoming Chen, Runyang Feng, Shuang Wu, Shouling Ji, Bailin Yang, and Xun Wang. Deep dual consecutive network for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 525–534, 2021. 2

[2] MMPose Contributors. OpenMMLab Pose Estimation Toolbox and Benchmark, 2020. 1

[3] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[4] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022. 1

[5] Renjie Zhang, Di Lin, Xin Wang, Ruonan Liu, Bin Sheng, George Baciu, CL Philip Chen, and Ping Li. Temporal-interim pose synthesis and distillation for dynamic human pose estimation. *IEEE Transactions on Neural Networks and Learning Systems*, 2025. 2