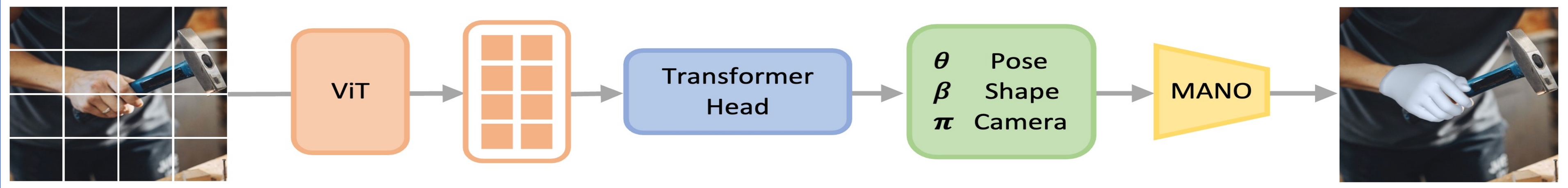
ExtPose: Robust & Coherent Pose Estimation by Extending ViTs

Glory Rongyu CHEN*, Li'an ZHUO*, Linlin YANG, Qi WANG, Liefeng BO, Bang ZHANG, & Angela YAO

Introduction

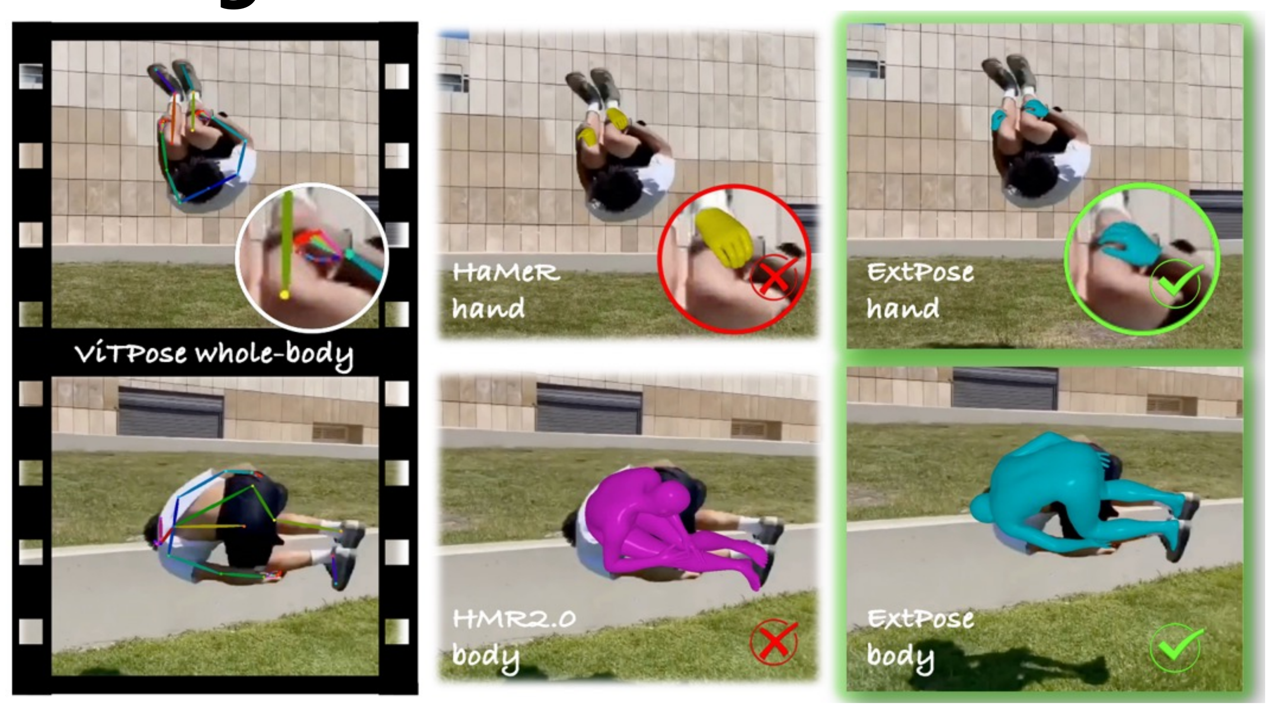
Task: 3D Human/Hand Pose Estimation (HPE)



•**Output:** 3D pose parameters of the Human & Hand Model. Projected to 2D for visualization

•**Architecture: Vision Transformers (ViTs)** working on image patches of size 16 x 16

Shortcomings



✗**Robustness: misalignment between 3D poses& images**, *e.g.* wrong orient. for complex motion in Col. 2 & occlusions

✗**Coherence: the ViT itself does not consider the temporal info** for videos, requiring an additional temporal module on top of frame features to alleviate jitter

Insights

Template Matching VS. Regression

•*Image alignment:* 2D HPE based on **template matching** is better than **regression**-based 3D HPE

•**Attention freely collects info on any relationship**, *e.g.* those **1.** between hands & bodies (spatial), **2.** between image & 2D pose modalities, & **3.** across frames (temporal)

Takeaways

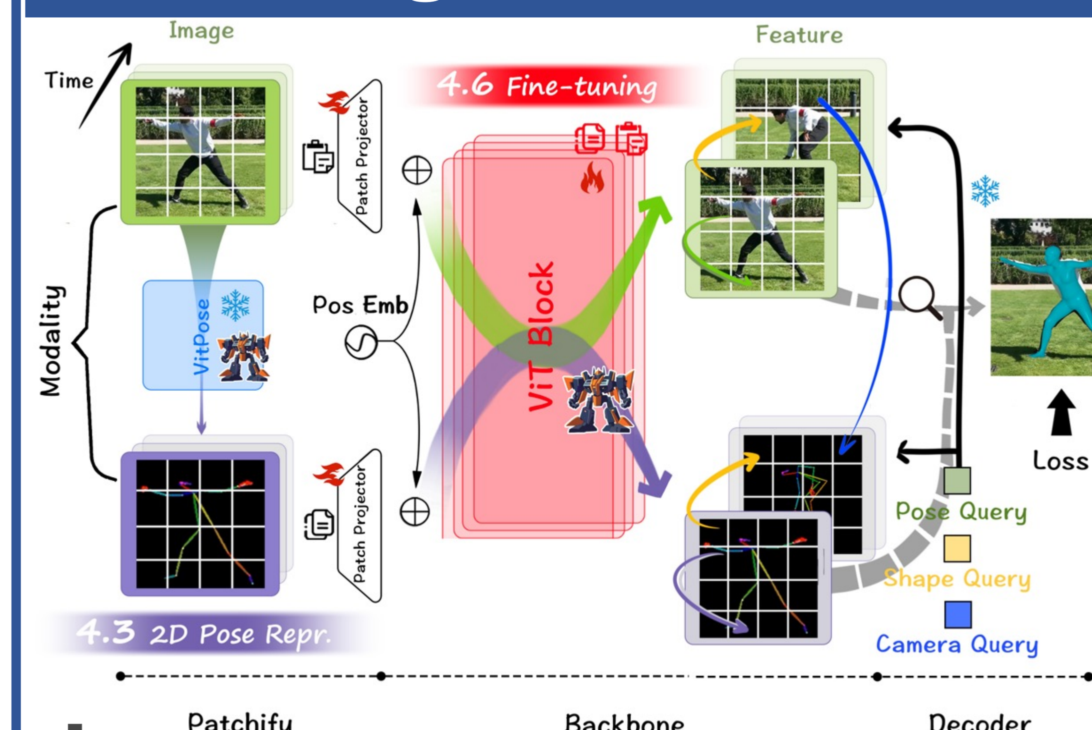
Unified HPE Framework with Modular Attention

•We extend the pose ViT into the first Video ViT pose estimator by re-programming the attention, which enhances robust & coherent features & can incorporate the info from multiple modalities, frames, & views, *etc.*

•**Logo:** complex **P**ose, attention (the cube) **E**xtension

•**Quotes:** *You can enjoy a grander sight, By climbing to a greater height* (Tang Peoms). *I.e.* leverage all available info incl. that from other parts, modalities, & frames

Extending Pose Vision Transformers

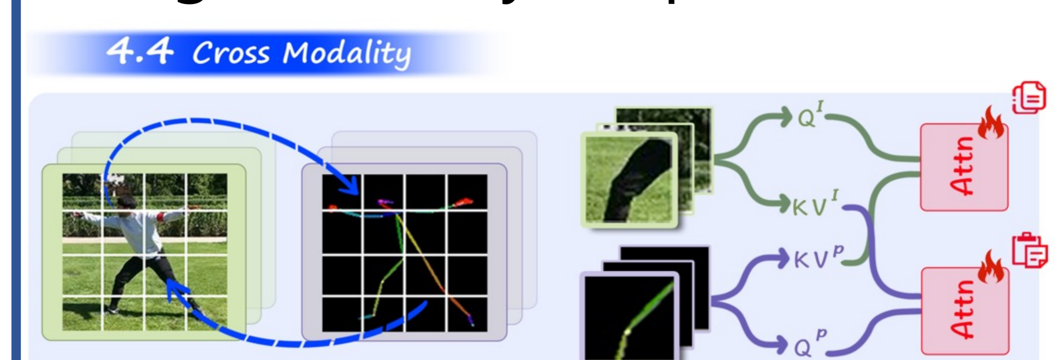


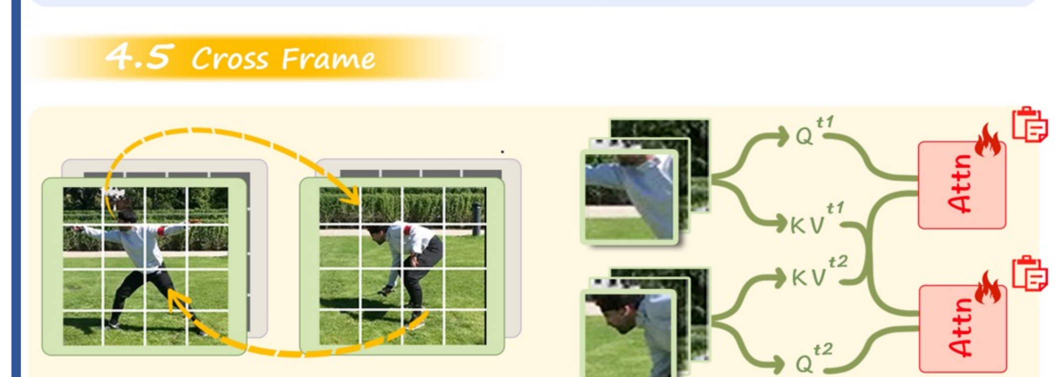
2D ViT Pose Form

Pose images (Col. 1) instead of 1D arrays are chosen for the same spatial layout as images & depicting joints & human configs

+Multi-Modal Pose ViTs

2D pose images & RGB images can be well processed & seamlessly fused by **one shared** vanilla ViTs with the **Multi-Modal Attention**, exploiting the layout but not being misled by 2D pose errors like Concat & ControlNet



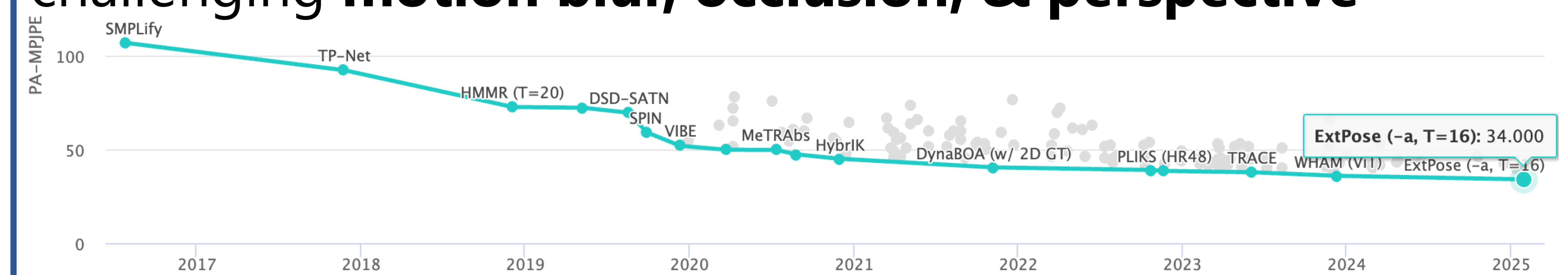


Video Pose ViTs

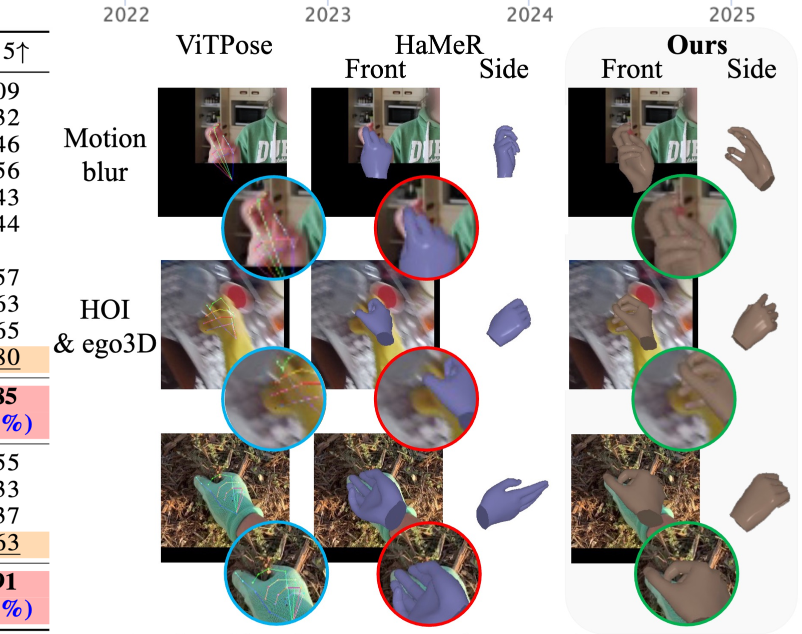
Attending & fusing features from multiple frames at **each layer** is more effective than just fine-tuning a **temporal head** on frame features

Extensive Experiments

SOTA on 5 human & hand datasets: 23% accuracy (PA-MPJPE) improvement on the 3DPW, ✓robust & coherent in challenging **motion blur, occlusion, & perspective**



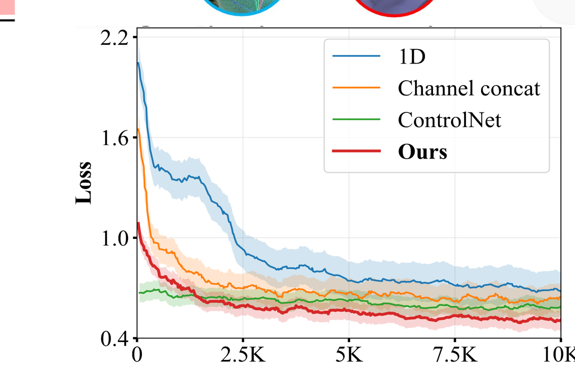
	Method	AUC$_J$↑	PJ↓	AUC$_V$↑	PV↓	F@5↑	F@15↑
Image	Pose2Mesh (ECCV'20)	0.754	12.5	0.749	12.7	0.441	0.909
	I2L-MeshNet (ECCV'20)	0.775	11.2	0.722	13.9	0.409	0.932
	METRO (CVPR'21)	0.792	10.4	0.779	11.1	0.484	0.946
	Liu *et al.* (CVPR'21)	0.803	9.9	0.810	9.5	0.528	0.956
	I2UV-HandNet (ICCV'21)	0.804	9.9	0.799	10.1	0.500	0.943
	ArtiBoost (CVPR'22)	0.773	11.4	0.782	10.9	0.488	0.944
	KeypointTrans (CVPR'22)	0.786	10.8	-	-	-	-
	MobRecon (CVPR'22)	-	9.2	-	9.4	0.538	0.957
	HandOccNet (CVPR'22)	0.819	9.1	0.819	8.8	0.564	0.963
	AMVUR (CVPR'23)	0.835	8.3	0.836	8.2	0.608	0.965
	HaMeR (CVPR'24)	0.846	7.7	0.841	7.9	0.635	0.980
	ExtPose ($T=1$)	**0.858** (+1.4%)	**7.0** (-9.1%)	**0.850** (+1.1%)	**7.5** (-5.1%)	**0.660** (+3.9%)	**0.985** (+0.5%)
Video	VIBE* (CVPR'20)	-	9.9	-	9.5	0.526	0.955
	TCMR* (CVPR'21)	-	11.4	-	10.9	0.463	0.933
	TempCLR* (3DV'22)	-	10.6	-	10.6	0.481	0.937
	Deformer (ICCV'23)	-	9.4	-	9.1	0.546	0.963
	ExtPose ($T=16$)	**0.863**	**6.9** (-26.6%)	**0.856**	**7.3** (-19.8%)	**0.667** (+22.2%)	**0.991** (+2.9%)



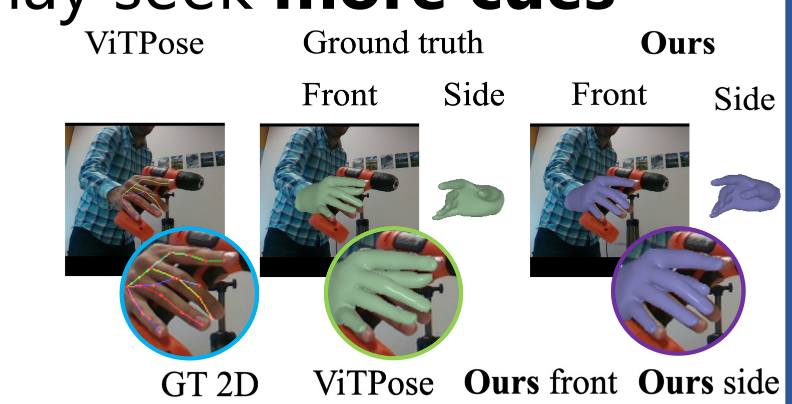
•**Ablations of 2D pose forms, modality fusions, & strategies**

Img	2D pose	PA-MPJPE↓	PA-MPVPE↓	F@5↑	F@15↑
	1D	6.5	6.6	0.724	0.983
	Heatmap	6.3	6.3	0.747	0.984
	Skel. image	6.2	6.3	0.742	0.985
✓		6.0	5.7	0.783	0.991
✓	Skel. image	**4.9**	**5.1**	**0.823**	**0.993**

Method	New Days @0.05	New Days @0.1	New Days @0.15	VISOR @0.05	VISOR @0.1	VISOR @0.15
HaMeR	48.0	78.0	88.8	43.0	76.9	89.3
Fusion						
Late fusion	50.5	82.4	92.5	52.5	87.1	95.6
Channel concat*	56.3	83.6	92.2	55.9	87.3	95.3
ControlNet*	55.6	83.5	92.3	57.7	87.5	95.5
Training						
From ViTPose	49.9	82.2	92.2	46.4	85.3	95.2
Only Q, K	50.0	81.9	92.3	49.1	85.3	95.1
1^{st} half	50.8	82.2	92.3	50.2	85.8	95.2
ExtPose	**59.6**	**84.8**	**92.7**	**61.1**	**88.5**	**95.6**



Thus, **robust to 2D pose errors;** yet, both branches **fail in extreme cases** where we may seek **more cues**



⚠**Multi-Modal Attention:** image features focus on the **keypoint** of 2D poses, while 2D pose features also examine the depth info in the RGB **background**