FinePOSE: Fine-Grained Prompt-Driven 3D Human Pose Estimation via Diffusion Models

Jinglin Xu, Yijie Guo, Yuxin Peng*
xujinglinlove@gmail.com; 2000012936@stu.pku.edu.cn; pengyuxin@pku.edu.cn

VALSE 2024 重庆 视觉与学习青年学者研讨会 VISION AND LEARNING SEMINAR

Motivations

- 3D Human Pose Estimation (3D HPE) is a task to estimate the 3D body joints and bones from 2D images or videos.
- 3DHPE is challenging because of its uncertainty (depth ambiguity) and complexity (complex human body structure).
- Most methods ignore the capability of coupling accessible texts and naturally feasible human knowledge, missing out on valuable implicit supervision to guide the 3D HPE task.
- Previous efforts often neglect fine-grained guidance hidden in different human body parts.

Contributions

- We propose FinePOSE, a new fine-grained part-aware prompt learning mechanism coupled with diffusion models.
- Our FinePOSE encodes multi-granularity information and establishes fine-grained communications between learnable part-aware prompts and poses.
- Extensive experiments illustrate that our approach obtains substantial improvements and achieves the state-of-the-art.

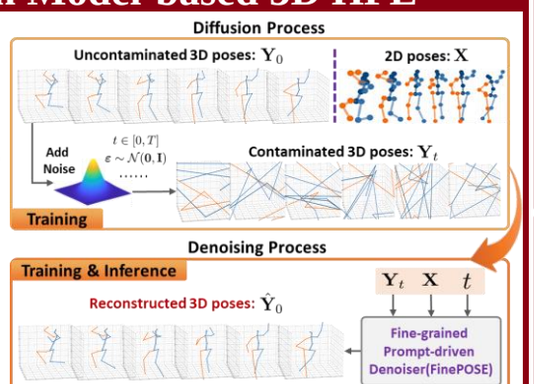
Diffusion Model-based 3D HPE

- Add Gaussian noise into Y_0

$$q(\mathbf{Y}_t \mid \mathbf{Y}_0) := \sqrt{\bar{\alpha}_t}\mathbf{Y}_0 + \epsilon\sqrt{1-\bar{\alpha}_t},$$

- Passing Y_t to D to get $\hat{Y}_0$
- Obtain Y_{t-1} at timestamp $t-1$

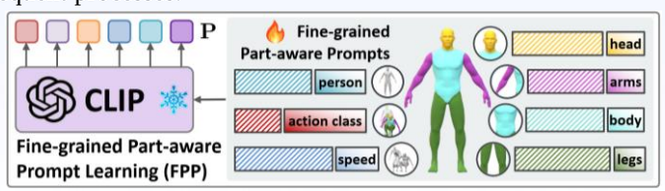
$$\mathbf{Y}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{\mathbf{Y}}_0 + \epsilon_t\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2} + \sigma_t\epsilon,$$



Method: FinePOSE

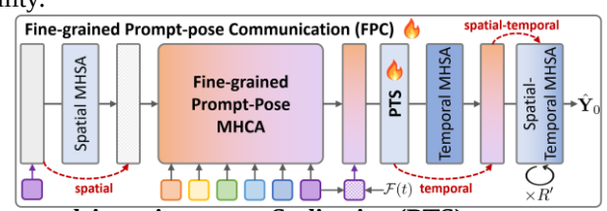
Fine-grained Part-aware Prompt Learning (FPP)

The FPP block encodes three kinds of information about the human pose, including action class, coarse- and fine-grained parts of humans like "person, head, body, arms, legs", and kinematic information "speed", and integrates them with pose features for serving subsequent processes.



Fine-grained Prompt-pose Communication (FPC)

The FPC block injects fine-grained part-aware prompt embedding into noise 3D poses to establish fine-grained communications between learnable part-aware prompts and poses for enhancing the denoising capability.



Prompt-driven timestamp Stylization (PTS)

The PTS block introduces the timestamp coupled with fine-grained part-aware prompt embedding into the denoising process to enhance its adaptability and refine the prediction at each noise level.

Project Page | MIPL's Website | MIPL's GitHub | Personal Home Page

Experiments

Method / MPJPE ↓	Human3.6M (DET)															
	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
TCN [29]	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
SRNet [51]	46.6	47.1	43.9	41.6	45.8	49.6	46.5	40.0	53.4	61.1	46.1	42.6	43.1	31.5	32.6	44.8
RIE [32]	40.8	44.5	41.4	42.7	46.3	55.6	41.8	41.9	53.7	60.8	45.0	41.5	44.8	30.8	31.9	44.3
Anatomy [6]	41.4	43.5	40.1	42.9	46.6	51.9	41.7	42.3	53.9	60.2	45.4	41.7	46.0	31.5	32.7	44.1
P-STMO [33]	38.9	42.7	40.4	41.1	45.6	49.7	40.9	39.9	55.5	59.4	44.9	42.2	42.7	29.4	29.4	42.8
MixSTE [52]	36.7	39.0	36.5	39.4	40.2	44.9	39.8	36.9	47.9	54.8	39.6	37.8	39.3	29.7	30.6	39.8
PoseFormerV2 [54]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	45.2
MHFormer [19]	39.2	43.1	40.1	40.9	44.9	51.2	40.6	41.3	53.5	60.3	43.7	41.1	43.8	29.8	30.6	43.0
Diffpose [10]	33.2	36.6	33.0	35.6	37.6	45.1	35.7	35.5	46.4	49.9	37.3	35.6	36.5	24.4	24.1	36.9
GLA-GCN [48]	41.3	44.3	40.8	41.8	45.9	54.1	42.1	41.5	57.8	62.9	45.0	42.8	45.9	29.4	29.9	44.4
ActionPrompt [55]	37.7	40.2	39.8	40.6	43.1	48.0	38.8	38.9	50.8	63.2	42.0	40.0	42.0	30.5	31.6	41.8
MotionBERT [59]	36.1	37.5	35.8	32.1	40.3	46.3	36.1	35.3	46.9	53.9	39.5	36.3	35.8	25.1	25.3	37.5
D3DP [34]	33.0	34.8	31.7	33.1	37.5	43.7	34.8	33.6	45.7	47.8	37.0	35.0	35.0	24.3	24.1	35.4
FinePOSE (Ours)	**31.4** (-1.6)	**31.5** (-3.3)	**28.8** (-2.9)	**29.7** (-2.4)	**34.3** (-3.2)	**36.5** (-7.2)	**29.2** (-5.6)	**30.0** (-3.6)	**42.0** (-3.7)	**42.5** (-5.3)	**33.3** (-3.7)	**31.9** (-3.1)	**31.4** (-3.6)	**22.6** (-1.7)	**22.7** (-1.4)	**31.9** (-3.5)

Method	N	MPI-INF-3DHP		
		PCK↑	AUC↑	MPJPE ↓
TCN [29]	81	86.0	51.9	84.0
Anatomy [6]	81	87.9	54.0	78.8
P-STMO [33]	81	97.9	75.8	32.2
MixSTE [52]	27	94.4	66.5	54.9
PoseFormerV2 [54]	81	97.9	78.8	27.8
MHFormer [19]	9	93.8	63.3	58.0
Diffpose [10]	81	98.0	75.9	29.1
GLA-GCN [48]	81	98.5	79.1	27.8
D3DP [34]	243	98.0	79.1	28.1
FinePOSE (Ours)	243	**98.7** (+0.2)	**79.7** (+0.6)	**26.8** (-1.0)

Method	N	Human3.6M (GT)		
		Detector	MPJPE ↓	P-MPJPE ↓
TCN [29]	243	GT	37.8	/
Anatomy [6]	243	GT	32.3	/
P-STMO [33]	243	GT	29.3	/
MixSTE [52]	243	GT	21.6	/
PoseFormerV2 [54]	243	GT	35.5	/
MHFormer [19]	351	GT	30.5	/
Diffpose [10]	243	GT	18.9	/
GLA-GCN [48]	243	GT	21.0	17.6
ActionPrompt [55]	243	GT	22.7	/
MotionBERT [59]	243	GT	16.9	/
D3DP [34]	243	GT	18.4	/
FinePOSE (Ours)	243	GT	**16.7** (-0.2)	**12.7** (-4.9)

Method / MPJPE ↓	EgoHumans			
	Tag.	Lego	Fenc.	Avg
D3DP [35]	30.7	29.0	46.6	35.4
FinePOSE (Ours)	**30.0** (-0.7)	**26.7** (-2.3)	**46.2** (-0.4)	**34.3** (-1.1)

MPJPE: mean per joint position error.
PCK: percentage of correct keypoint.
AUC: area under curve.

Visualization

The 3D poses predicted by FinePOSE match better with ground-truth 3D poses than other methods.