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

Jinglin Xu, Yijie Guo, Yuxin Peng

by

CVPR 2024

2024.11.26

Index

Introduction

- 3D Human Pose Estimation(HPE)를 하는 데에는, 일반적으로 아래의 두가지 과정을 거침
  - 이미지 또는 비디오에서 2D keypoint를 추출
  - 2D keypoint로부터 3D Pose를 mapping
- 본 논문에서는 두 번째 과정에 집중함

Introduction

- 3D Pose Estimation은 아래와 같은 세 문제가 존재함
  - 2D에서 depth정보가 없기 때문에 3D로 바꾸는 데 모호함이 생김
  - 유연한 인체와 복잡한 관절 간의 관계로 self occlusion 등이 생길 수 있음
  - 현존하는 데이터 세트는 action class가 많지 않기에 overfitting이 발생하거나, 복잡한 action으로 뻗어나가기 어려움

Introduction

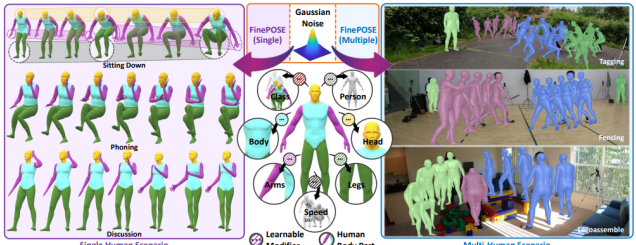


Figure 1. **Illustration of Fine-grained Prompt-driven Denoiser (FinePOSE).** FinePOSE, the proposed diffusion model-based 3D human pose estimation approach, enables multi-granularity manipulation controlled by learnable modifiers (e.g., "action class", coarse- and fine-grained human body parts including "person, head, body, arms, legs", and kinematic information "speed"), boosting motion reconstruction for single human and multi-human scenarios.

- 이러한 문제를 해결하기 위해, 입력 정보를 개선함
  1. 인간의 action class
  2. 속도
  3. 각 부위가 어떻게 움직이는지

Index

Approach

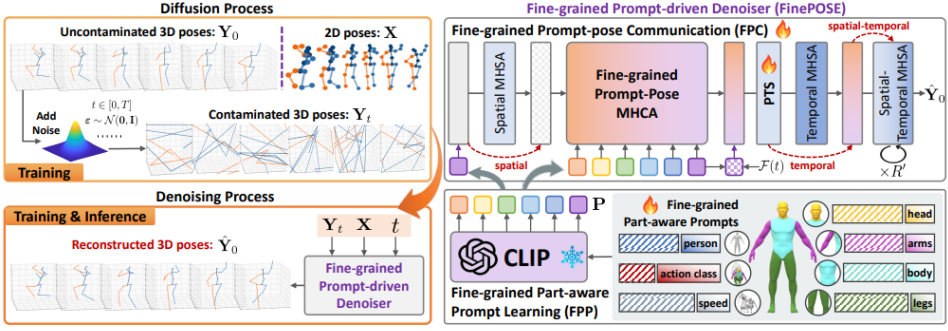
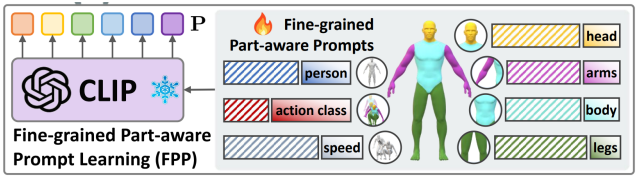


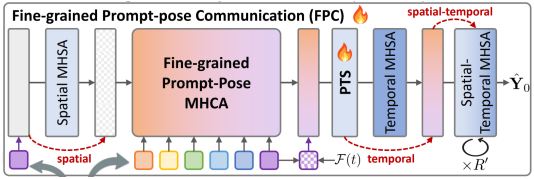
Figure 2. **The architecture of the proposed FinePOSE.** In the diffusion process, Gaussian noise is gradually added to the ground-truth 3D poses $\mathbf{Y}_0$, generating the noisy 3D poses $\mathbf{Y}_t$ for the timestamp t. In the denoising process, $\mathbf{Y}_t$, $\mathbf{X}$ and t are fed to fine-grained prompt-driven denoiser $\mathcal{D}$ to reconstruct pure 3D poses $\hat{\mathbf{Y}}_0$. $\mathcal{D}$ is composed of a Fine-grained Part-aware Prompt learning (FPP) block, a Fine-grained Prompt-pose Communication (FPC) block, and a Prompt-driven Timestamp Stylization (PTS) block, where FPP provides more precise guidance for all human part movements, FPC establishes fine-grained communications between learnable prompts and poses for enhancing the denoising capability, and PTS integrates learned prompt embedding and current timestamp for refining the prediction at each noise level.

Approach



- Fine-grained Part-aware Prompt Learning(FPP)
- noise가 들어간 3D pose에서 원본을 재생성하기 위해 2D keypoint(X), timestamp(t), Embedding(P)가 필요함
- FPP 블록은 action class, 속도, 신체의 움직임 정보를 Embedding 공간 P에 인코딩하는 것을 학습함

Approach



- Fine-grained Prompt-pose Communication (FPC)
- 먼저, noise 정보 Y_t와 X, t, P를 합침

$$Z_t = Concat(Y_t, X) + P[L] + \mathcal{F}(t)$$

- Z_t를 Spatial transformer(Multi-Head Self Attention)에 넣어 Z_t^s를 얻음
  - 한 frame내 관절 간의 세밀한 관계에 집중 가능

- 추가로, Cross Attention 레이어를 추가하여 P의 정보를 완벽하게 주입한 Z_t^{sp}를 얻음

$$Q = W_Q Z_t^s, K = W_K P, V = W_V P$$

Cross Attention $A = softmax(Q \otimes K^\top / \sqrt{d})$

$$Z_t^{sp} = A \otimes V, \tilde{Z}_t^{sp} = \mathcal{P}(Z_t^{sp}), \mathcal{P}\text{는 PTS 블럭}$$

- Prompt-driven timestamp Stylization (PTS)
  - PTS 블럭은 P와 timestamp t를 선형으로 결합하는 블럭
  - 복원 중간에 timestamp 정보를 도입하여 예측을 세분화
- 이후 pose의 frame 간 관계를 모델링하기 위해 Temporal transformer에 넣고, 이를 한번 더 Spatial-Temporal transformer를 통해 3D Pose $\hat{Y}_0$을 예측

Index

Result

Method	N	Human3.6M (DET)			Human3.6M (GT)			Year
		Detector	MPJPE ↓	P-MPJPE ↓	Detector	MPJPE ↓	P-MPJPE ↓	
TCN [29]	243	CPN	46.8	36.5	GT	37.8	/	CVPR'19
Anatomy [6]	243	CPN	44.1	35.0	GT	32.3	/	CSVT'21
P-STMO [33]	243	CPN	42.8	34.4	GT	29.3	/	ECCV'22
MixSTE [52]	243	HRNet	39.8	30.6	GT	21.6	/	CVPR'22
PoseFormerV2 [54]	243	CPN	45.2	35.6	GT	35.5	/	CVPR'23
MHFormer [19]	351	CPN	43.0	34.4	GT	30.5	/	CVPR'22
Diffpose [10]	243	CPN	36.9	<u>28.7</u>	GT	18.9	/	CVPR'23
GLA-GCN [48]	243	CPN	44.4	34.8	GT	21.0	17.6	ICCV'23
ActionPrompt [55]	243	CPN	41.8	29.5	GT	22.7	/	ICME'23
MotionBERT [59]	243	SH	37.5	/	GT	<u>16.9</u>	/	ICCV'23
D3DP [34]	243	CPN	<u>35.4</u>	<u>28.7</u>	GT	18.4	/	ICCV'23
FinePOSE (Ours)	243	CPN	**31.9** (-3.5)	**25.0** (-3.7)	GT	**16.7** (-0.2)	**12.7** (-4.9)	

Table 1. **Quantitative comparison with the state-of-the-art 3D human pose estimation methods on the Human3.6M dataset.** N: the number of input frames. CPN, HRNet, SH: using CPN [7], HRNet [39], and SH [24] as the 2D keypoint detectors to generate the inputs. GT: using the ground truth 2D keypoints as inputs. The best and second-best results are highlighted in **bold** and <u>underlined</u> formats.

Result

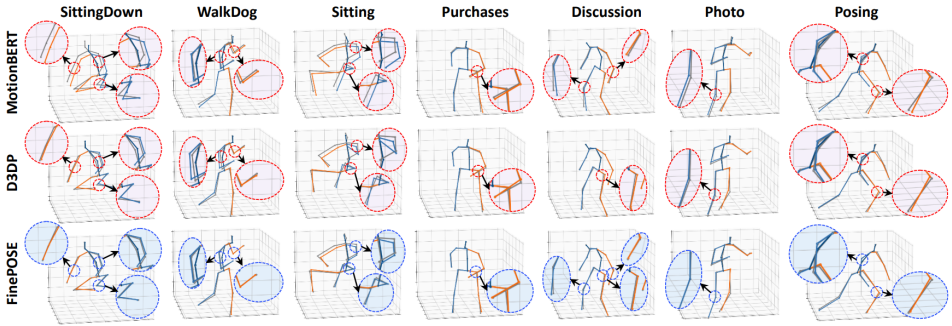


Figure 3. **Qualitative comparisons of our FinePOSE with MotionBERT [59] and D3DP [34] on Human3.6M.** The gray skeleton is the ground-truth 3D pose. The blue skeleton represents the prediction of the human left part, and the orange indicates the human right part. The red dashed line represents the incorrect regions of the compared methods, and the blue dashed line indicates the counterparts of FinePOSE.

Result

- FPP 블럭을 통해 신체 부위마다 정확한 예상을 할 수 있음
- FPC 블럭에서 prompt와 2D keypoint간 세밀한 관계를 구축하여 성능을 향상
- PTS 블럭을 통해 timestamp 정보를 도입하여 예측을 세분화
- 이 세 가지를 통해 당시 SOTA를 넘을 수 있었음
- Multi-human scenario로도 확장할 수 있으나, 이를 위해 설계된 것이 아니기 때문에 계산 비용이 높아지는 것이 Limitation