PADS: Plug-and-Play 3D Human Pose Analysis via Diffusion Generative Modeling

Haorui Ji, Hongdong Li
The Australian National University

Introduction

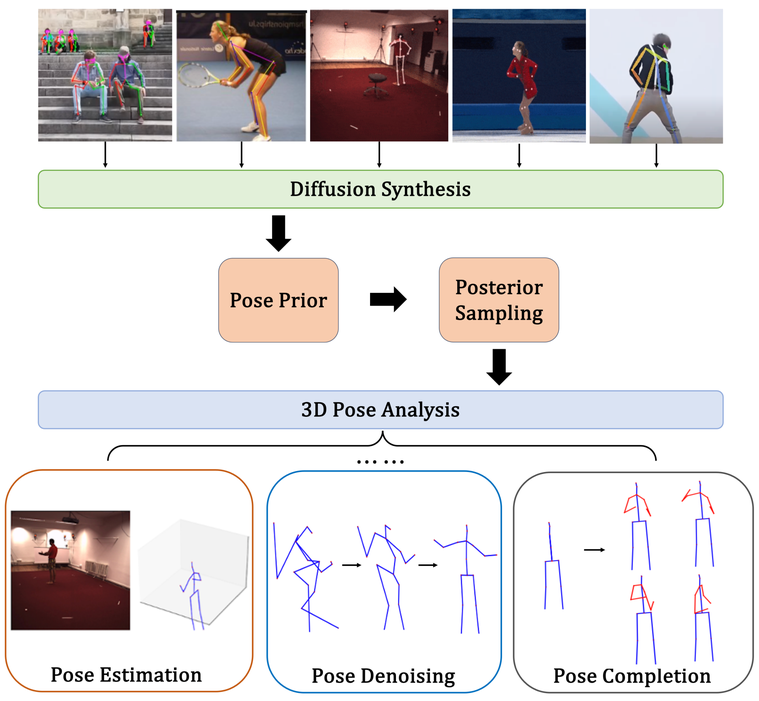
Goals and Challenges:

3D human analysis: reconstruct 3D human pose under different conditions

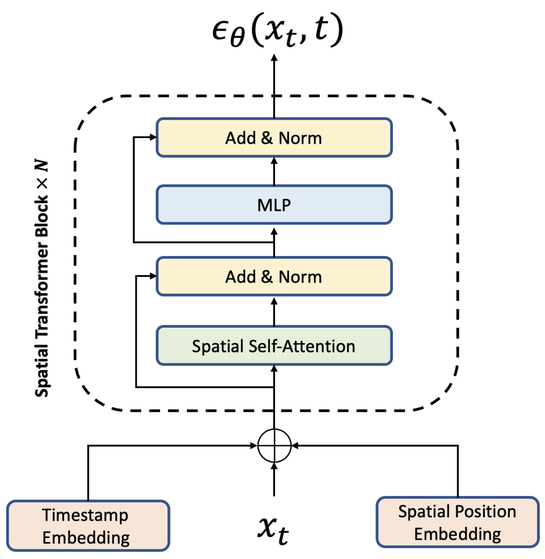
- Limited generalization across task boundaries.
- Performance depends heavily on access to paired data supervision.

Key Contributions:

- A unified generative framework that formulates diverse 3D human pose analysis tasks as inverse problems.
- Plug-and-play inference paradigm capable of handling diverse conditions without the need for retraining or architectural modification.



Diffusion Pose Prior



$$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) = -\frac{1}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t), \quad \bar{\alpha}_t = \prod_{i=1}^{t} (1-\beta_t)$$

Diffusion Posterior Sampling

Algorithm 1 PADS for 3D Human Pose Estimation

Require:
2D pose measurement P_{2d}; Camera intrinsic K; Pelvis root trajectory T;
Pose prior $\epsilon_\theta(\mathbf{x}_t, t)$; Number of diffusion sampling iterations N;
Guidance scale $\{\rho_t\}_{t=1}^N$; Noise variance $\{\sigma_t\}_{t=1}^N$;

$\widetilde{P}_{init} \leftarrow \frac{K^{-1}P_{2d}}{\|K^{-1}P_{2d}\|_2}\|T\|_2$ ▷ Pose initialization
$P_{init} \leftarrow \widetilde{P}_{init} - \widetilde{P}_{init}^{pelvis}$
$\mathbf{x}_N \leftarrow P_{init}$, $\mathbf{y} \leftarrow P_{2d}$
for $t \leftarrow N$ to 1 **do** ▷ Iterative reverse sampling
  $\hat{\mathbf{x}}_0 \leftarrow \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t, t))$
  $\mathbf{z} \sim N(0, \mathbf{I})$
  $\mathbf{x}'_{t-1} \leftarrow \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{\mathbf{x}}_0 + \sigma_t \mathbf{z}$
  $\mathbf{x}_{t-1} \leftarrow \mathbf{x}'_{t-1} - \rho_t \nabla_{\mathbf{x}_t} \|\mathbf{y} - proj(\hat{\mathbf{x}}_0)\|_2^2$
end for
return $\mathbf{x}_0$

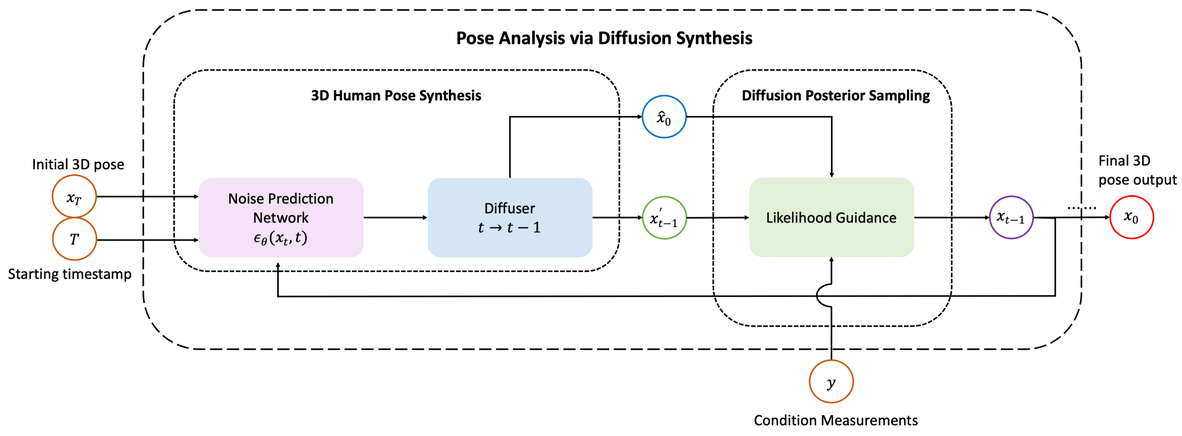
Approximation quantified with the Jensen gap

$$\nabla_{\mathbf{x}_t} \log p(\mathbf{y}|\mathbf{x}_t) \simeq \nabla_{\mathbf{x}_t} \log p(\mathbf{y}|\hat{\mathbf{x}}_0) = -\frac{1}{\sigma^2} \nabla_{\mathbf{x}_t} \|\mathbf{y} - f(\hat{\mathbf{x}}_0)\|_2^2$$

Combining with the learned pose prior yields,

$$\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{y}) \simeq -\frac{1}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) - \rho \nabla_{\mathbf{x}_t} \|\mathbf{y} - f(\hat{\mathbf{x}}_0)\|_2^2$$

Pose Analysis by Diffusion Synthesis



Conditional sampling via Bayes' theorem decomposition:

$$\mathrm{d}\mathbf{x} = [-\frac{\beta(t)}{2}\mathbf{x} - \beta(t)\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y})]\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}\mathbf{w}$$
$$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y}) = \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p_t(\mathbf{y}|\mathbf{x}_t)$$

Two-stage pipeline:

- **Training**: unconditional 3D human pose synthesis to model the underlying kinematic constraints
- **Inference**: utilize the learned prior along with diffusion posterior sampling scheme to inject likelihood guidance into the sampling process.

Experiments & Results

Pose Estimation

Mode	Method	GT MPJPE ↓	GT PA-MPJPE ↓	DT MPJPE ↓	DT PA-MPJPE ↓
Lrn.	Yu *et al.* [42]	85.3	42.0	92.4	54.3
	PAUL [40]	88.3	-	132.5	-
	MHR-Net [44]	72.6	-	-	-
	ElePose [39]	64.0	36.7	-	-
Opt.	SMPL [2]	82.3	-	-	-
	Song *et al.* [34]	-	56.4	-	-
	ZeDO [16] ($H=1$)	-	-	65.7	49.0
	ZeDO [16] ($H=10$)	-	-	57.1	45.1
	PADS	**41.5**	**33.1**	**54.8**	**44.9**

Method	GT-MPJPE ↓	DT-MPJPE ↓
DiffPose (CVPR2023) [10]	**31.6**	49.7
DiffPose (IROS2023) [5]	37.3	49.4
D3DP [31]	34.24	**44.3**
DiffHPE [30]	-	51.2
PoseFormerV2 [46]	37.3	47.5
MotionBERT [48]	37.5	-
SPIN [18]	41.1	-
HybrIK-X [19]	47.0	-
PADS	41.5	54.8

Pose Generation

Sample source	APD ↑
AMASS[23]	15.44
VPoser[26]	10.75
PoseNDF[38]	**18.75**
DPoser[22]	14.28
PADS	15.81

Pose Completion

Masked Body Parts	MPJPE ↓
Right Leg	17.1
Left Leg	18.8
Right Arm	32.6
Left Arm	30.3
Two Legs	63.3
Two Arms	72.8
Spine	4.5

Pose Denoising

Noise Type	Intensity	Before↓	After↓
Gaussian	$\mathcal{N}(0, 0.25)$	75.3	51.6
	$\mathcal{N}(0, 0.5)$	150.5	63.8
	$\mathcal{N}(0, 1.0)$	301.2	88.7
Uniform	$\mathcal{U}(-0.25, 0.25)$	45.2	43.1
	$\mathcal{U}(-0.5, 0.5)$	90.4	52.9
	$\mathcal{U}(-1.0, 1.0)$	180.9	69.6
Poisson	$\mathcal{P}(0.25)$	152.3	90.1
Salt and Pepper	$p = 0.25$	149.1	108.9