DiffPose: Toward More Reliable 3D Pose Estimation

Jia Gong[1,*], Lin Geng Foo[1,*], Zhipeng Fan[2], Qiuhong Ke[3], Hossein Rahmani[4], Jun Liu[1]
1 Singapore University of Technology and Design, Singapore; 2 New York University, United States;
3 Monash University, Australia; 4 Lancaster University, United Kingdom; * Equal contribution

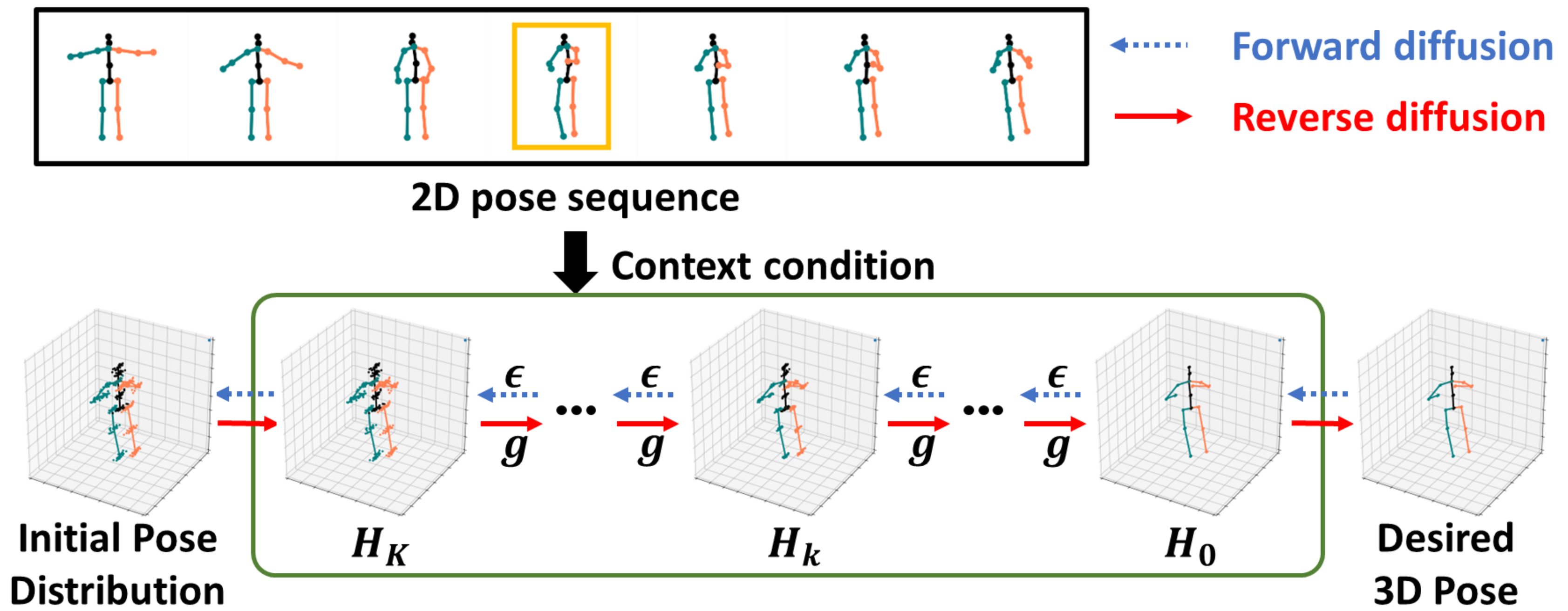
SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN · NEW YORK UNIVERSITY · MONASH University · Lancaster University

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Motivation and Overview

- **Goal:** Handle the uncertainty in 3D pose estimation (due to depth ambiguity and potential occlusion).
- **Motivation:** Inspired by the strong capability of diffusion models to generate high-quality samples from random noise, we tackle 3D pose estimation, which involves uncertainty and indeterminacy, with diffusion models.
- **Our method:** <u>We formulate 3D pose estimation as a reverse diffusion process</u> and propose various designs, including the initialization of 3D pose distribution, <u>a GMM-based forward diffusion process</u> and <u>a conditional reverse diffusion process</u>.

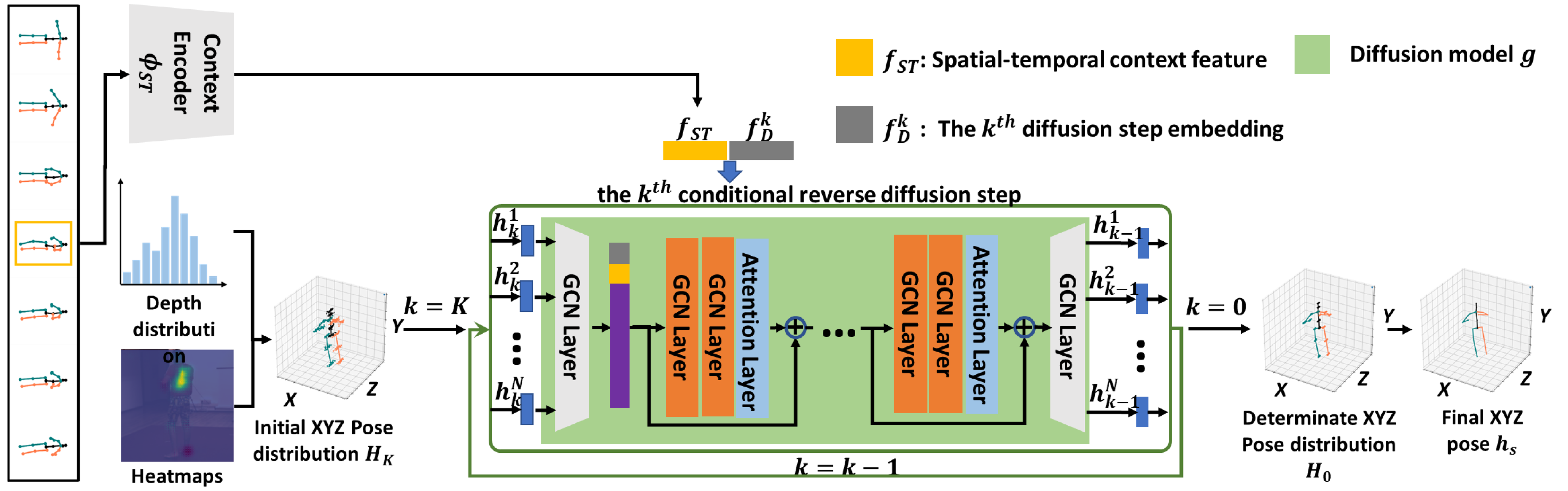
Pose Diffusion Process



In the forward process, we gradually diffuse a "ground truth" 3D pose distribution H_0 with low indeterminacy towards a 3D pose distribution with high uncertainty adding noise ϵ at every step.

In the reverse process, we first initialize the indeterminate 3D pose distribution $\boldsymbol{H_K}$ from the input. Then, during the reverse process, we use the diffusion model g, conditioned on the context information from 2D pose sequence, to progressively transform $\boldsymbol{H_K}$ into a 3D pose distribution $\boldsymbol{H_0}$ with low indeterminacy.

Our DiffPose Framework



Initializing 3D Pose Distribution H_K:

We initialize the pose distribution H_K using **heatmaps** derived from a 2D pose detector and depth distributions that can either be computed from the training set or predicted by the Context Encoder.

Forward Diffusion Process:

During model training, we utilize forward diffusion process to generate the indeterminate 3D pose distributions that eventually (after K steps) resemble $\boldsymbol{H_K}$, we add noise to the ground truth 3D pose distribution $\boldsymbol{H_0}$. The noise is modeled by a **Gaussian Mixture Model (GMM)** that characterizes the uncertainty distribution H_K, which is obtained by fitting H_K with the EM algorithm.

Reverse Diffusion Process:

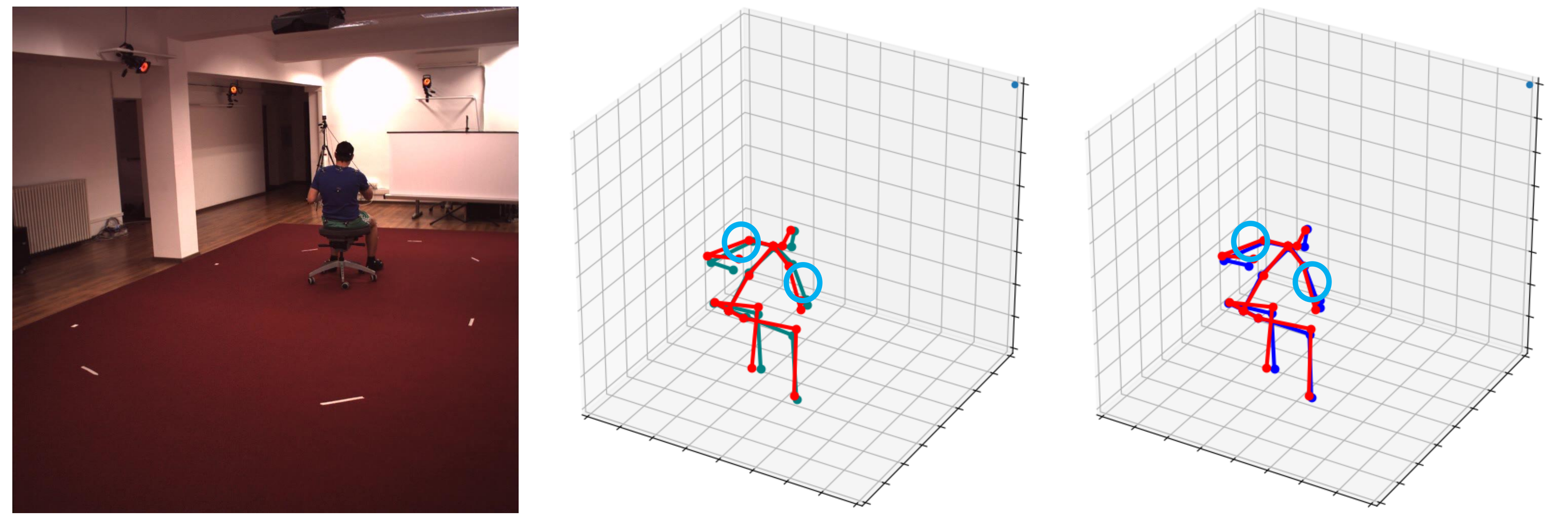
The reverse diffusion process is conditioned on context information (extracted via a Context Encoder) from the input video or frame in order to better leverage the spatial-temporal relationship between frames and joints. Then, to effectively use the context information and perform the progressive denoising to obtain accurate 3D poses, we design a GCN-based diffusion model g.

[1] Zhang, Jinlu, et al. "Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
[2] Zhao, Weixi, Weiqiang Wang, and Yunjie Tian. "GraFormer: Graph-oriented transformer for 3D pose estimation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

Experimental Results

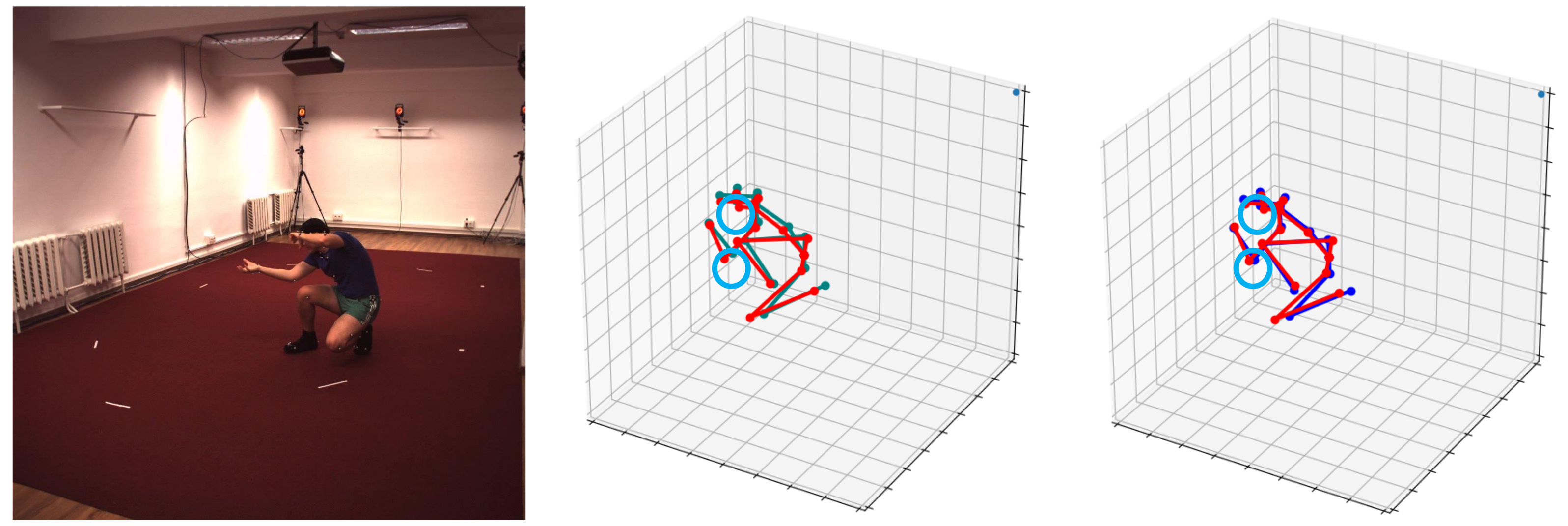
Video-based results:

Image | MixSTE [1] | DiffPose



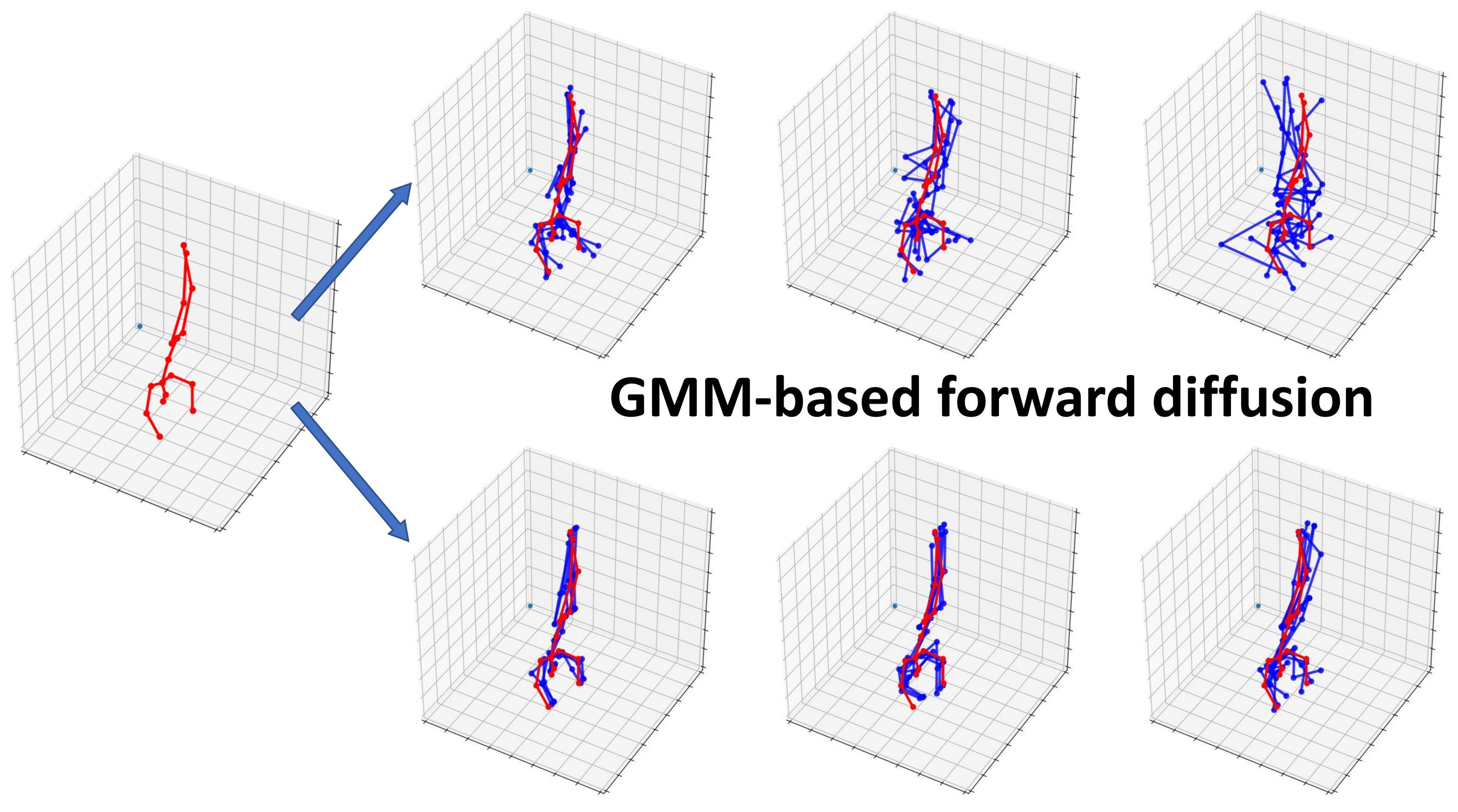
Frame-based results:

Image | Graformer [2] | DiffPose



Qualitative comparisons:

Standard forward diffusion



GMM-based forward diffusion

k=5 | k=10 | k=15