HandDiff: 3D Hand Pose Estimation with Diffusion on Image-Point Cloud

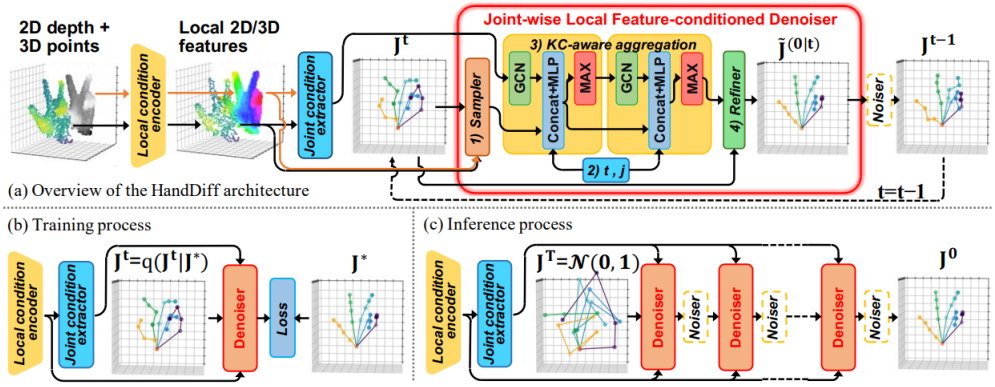
Wencan Cheng, Hao Tang, Luc Van Gool, Jong Hwan Ko

Accepted for CVPR 2024 (Paper) (GitHub)

This paper introduces a new 3d hand pose estimation model based on diffusion models. The model uses depth images and hand-shaped 3D point clouds, samples a random set of joint positions from a normal distribution and then, iteratively denoises the hand poses and estimates joint locations.

Motivation

Self-occlusions is a common problem for hand pose as finger joints are often occluded. Also, previous 3D diffusion models is that they rely on global features and overlook local details. Another issue of 3D diffusion models is that they are permutation-equivariant which limits their ability to distinguish between joints.



Method

1. **Joint-wise Condition Extraction**

   The ConvNeXt-based encoder processes the depth image to generate a 2D local visual feature map and a 2D global vector. The PointNet++ encoder processes the 3D point cloud to gemerate 3D local geometric features and a 3D global vector. The global vectors are concatenated to form a single global representation. The global representation is passed through three layers of Bias-Induced Layers (BIL) to generate joint-wise embeddings (joint-wise conditions).

2. **Joint-wise Local Feature-conditioned Denoiser**

   The denoiser starts with a set of joint coordinates that are randomly sampled from a Gaussian distribution. At each time step, the denoiser samples local features around each noisy joint in both 2D and 3D spaces, encodes the joint index and timestep into embeddings, aggregates local features with joint-wise conditions using a kinematic-aware GCN block, and outputs the refined joint coordinates for the next step.

Method	Mean joint error (mm)			Input
	ICVL	MSRA	NYU	
DeepPrior++ [31]	8.1	9.5	12.24	D
Pose-Ren [6]	6.79	8.65	11.81	D
DenseReg [52]	7.3	7.2	10.2	D
CrossInfoNet [11]	6.73	7.86	10.08	D
JGR-P2O [12]	6.02	7.55	8.29	D
SSRN [37]	6.01	7.05	7.37	D
PHG [36]	5.97	6.94	7.39	D
VVS [7] †	6.22	-	7.79	D
3DCNN [14]	-	9.6	14.1	V
SHPR-Net [5]	7.22	7.76	10.78	P
HandPointNet [15]	6.94	8.5	10.54	P
Point-to-Point [16]	6.3	7.7	9.10	P
V2V [29]	6.28	7.59	8.42	V
HandFolding [9]	5.95	7.34	8.58	P
IPNet [38]	5.76	6.92	**7.17**	D+P
HandDiff (Ours)	**5.72**	**6.53**	7.38	D+P

Limitations

1. The model's performance plateaus after 5 timesteps which limits further optimization.
2. The model is unable to handle scenarios with interacting hands and focuses on single-hand samples, as stated in the conclusion.