SOTA Methods in 3D Hand and Body Pose from CVPR 2024 and ECCV 2024

HandDAGT

- Conference: ECCV 2024
- Authors: Wencan Cheng, Eunji Kim, Jong Hwan Ko
- Hand or Body: Hand
- Motivation: Existing graph methods in 3D hand pose estimation use static graphs which are unable to capture dynamic kinematic relations between joints in occluded scenarios. Also, self-occlusion and hand-object occlusion are major challenges for current methods.
- Method: The 3D hand point cloud and depth map are used to generate keypoint embeddings and determine a 3d point patch for each joint. The embeddings and patches are passed as queries and keys to an adaptive graph transformer, which dynamically balances local attention for visible keypoints and kinematic attention for occluded ones.
- Limitations: The model can't process interacting hands and focuses on single-hand samples. Also, this method requires redundant computation due to the 2D feature extraction. They state that possible solutions include bidirectional learning and developing a lightweight 2D model.
- Paper: https://arxiv.org/pdf/2407.20542
- GitHub: https://github.com/cwc1260/HandDAGT

HandDiff

- Conference: CVPR 2024
- Authors: Wencan Cheng, Hao Tang, Luc Van Gool, Jong Hwan Ko
- Hand or Body: Hand
- Motivation: Depth image-based methods either use 2D CNN-based approaches, which do not accurately capture the 3D structure, or 3D CNN-based approaches, which require large amounts of memory and computation. More recently, PointNet-based methods that process the 3D point cloud use a sparse point cloud to reduce computation at the cost of performance. Previous diffusion-based models rely on global features and overlook local details. Also, they are permutation-equivariant, which limits their ability to distinguish between joints.
- Method: The depth map and 3D point cloud are processed separately to generate 2D and 3D features. They are then concatenated and passed through three layers of BIL to generate joint-wise embeddings. The denoiser starts with a randomly sampled set of joint coordinates from a Gaussian distribution, and then for each time step out of a set amount, the denoiser samples local features around each joint and refines the joint coordinates using a kinematic-aware GCN.
- Limitations: The model can't process interacting hands and focuses on single-hand samples. Also, the denoiser's performance plateaus after around 5 timesteps.
- Paper: https://arxiv.org/pdf/2404.03159
- GitHub: https://github.com/cwc1260/HandDiff

HOISDF

- Conference: CVPR 2024
- Authors: Haozhe Qi, Chen Zhao, Mathieu Salzmann, Alexander Mathis
- Hand or Body: Hand
- Motivation: Direct lifting methods do not utilize 3D intermediate representations and rely on the network to learn the mapping from 2D image to 3D pose. Coarse-to-fine methods make initial predictions and then refine them with explicit intermediate representations, such as hand joints or vertices, but require more computation and is not as precise as implicit intermediate representations, such as a signed distance field as a continuous function. Also, current signed distance field applications mainly use them to directly reconstruct 3D meshes instead of as an intermediate representation.
- Method: This method estiamtes signed distance field for both the hand and object to build the shape of each. The 3D space is split into many bins, each with a query point, and the ones on the surface, which have the lowest distance to the surface, are then used to extract hand and object features. These features are enhanced with multi head self-attention layers and then used to compute the 3D hand joints and mesh, as well as the object rotation and translation vectors. Each joint's 3D position is fine-tuned with the distance vectors from each nearby query point.
- Limitations: The hand and object meshes might intersect with each other in severely occluded scenarios. Also, this model was only trained on objects that are bigger than the hand, which are much less occluded than smaller objects.