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

Wencan Cheng[1] Hao Tang[2,3] Luc Van Gool[2,4] Jong Hwan Ko[5]
[1]Department of Artificial Intelligence, Sungkyunkwan University
[2]CVL, ETH Zurich [3]Carnegie Mellon University [4]INSAIT, Sofia Un. St. Kliment Ohridski
[5]College of Information and Communication Engineering, Sungkyunkwan University
{cwc1260, jhko}@skku.edu, {hao.tang, vangool}@vision.ee.ethz.ch

Abstract

Extracting keypoint locations from input hand frames, known as 3D hand pose estimation, is a critical task in various human-computer interaction applications. Essentially, the 3D hand pose estimation can be regarded as a 3D point subset generative problem conditioned on input frames. Thanks to the recent significant progress on diffusion-based generative models, hand pose estimation can also benefit from the diffusion model to estimate keypoint locations with high quality. However, directly deploying the existing diffusion models to solve hand pose estimation is non-trivial, since they cannot achieve the complex permutation mapping and precise localization. Based on this motivation, this paper proposes HandDiff, a diffusion-based hand pose estimation model that iteratively denoises accurate hand pose conditioned on hand-shaped image-point clouds. In order to recover keypoint permutation and accurate location, we further introduce joint-wise condition and local detail condition. Experimental results demonstrate that the proposed HandDiff significantly outperforms the existing approaches on four challenging hand pose benchmark datasets. Codes and pre-trained models are publicly available at https://github.com/cwc1260/HandDiff.

1. Introduction

3D hand pose estimation (HPE), which involves estimating the 3D positions of hand keypoints, provides a fundamental conprehension of human hand motion. Therefore, it is essential to facilitate more natural and intuitive interactions between humans and machines and is applicable to various human-computer interaction applications including robotics, gaming, and augmented/virtual reality. In recent years, significant progress has been made in the field of 3D hand pose estimation by applying deep learning techniques and using low-cost depth cameras.

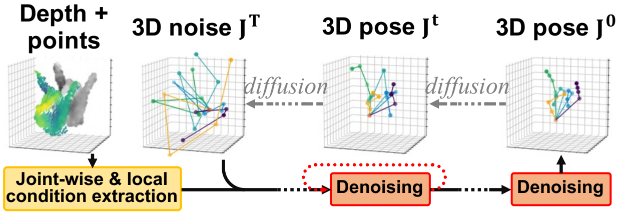


Figure 1. Illustration of the hand pose diffusion concept. The model extracts features from input depth images and corresponding point clouds as joint-wise and local conditions to guide the iterative denoising process that recovers accurate hand poses from diffused noisy pose distributions.

Recent developments in 3D Hand Pose Estimation (HPE) based on deep learning [5, 6, 9, 11, 12, 15, 16, 18, 23, 35, 36, 38] are primarily divided into two core approaches: regression and detection. While these straightforward solutions have shown notable effectiveness and computational efficiency, these deterministic methods impose limitations on handling ill-posed uncertain cases such as self-occlusions and hand-object occlusions, which are prevalent in real-world hand recognition scenarios. Therefore, in order to ensure the reliability of the estimation, it is imperative to accurately model the uncertainty.

Providentially, a revolutionary approach known as the Diffusion Model (DM) [19, 27, 43] has demonstrated remarkable performance in processing uncertainty through modeling of probabilistic distributions. The DM has also exhibited superiority in 3D generative applications including unseen 3D point cloud generation [27, 50] and uncertain parts completion [28, 54]. Therefore, 3D DM can be deployed in 3D HPE as in these 3D conditional generation problems, as both 3D HPE and 3D conditional generation aim to generate a set of keypoints based on a specific condition. More importantly, 3D DM can resolve the ill-posed uncertainty of occlusions by learning the probabilistic distribution of the keypoints. Based on this inspiration, we apply the diffusion model in generating hand keypoint locations conditioned on the hand depth image/point cloud input, as illustrated in Figure 1. To the best of our knowledge, our work is the first attempt to deploy diffusion models in the hand pose estimation task.

While it is theoretically possible to tackle the HPE task using DM models, the direct application of existing 3D DMs [27, 50] to hand pose estimation tasks is still limited. One of the significant limitations of current 3D DMs is their reliance on a global latent condition, which overlooks crucial local detail information needed for accurate estimation of joint locations. Furthermore, the permutation-equivariant nature of 3D DMs limits their ability to distinguish between joints under simple global conditions, affecting their precision in aligning noisy points with specific target joints.

To fully exploit the potential of the diffusion model in hand pose estimation, we propose HandDiff, a novel approach that incrementally refines the noise distribution to accurately derive a 3D hand pose from multi-modal inputs, including depth images and point clouds. To address inherent limitations in 3D DMs, our model incorporates a joint-wise denoising mechanism that individually denoises various joints during estimation. Concretely, the proposed model first introduces a joint-wise condition generation module that samples features for each individual joints from both depth image and point cloud. Furthermore, it features a novel local feature-conditioned denoising module, which is the key component to perform the reverse diffusion process. It operates under joint-specific conditions and utilizes local features gathered around the noisy input joint locations. In addition, we propose a novel kinematic correspondence-aware layer that integrates with a graph convolutional operation in order to capture the kinematic relationship between hand joints.

We evaluate HandDiff on four challenging benchmarks, including single-hand ICVL [47], MSRA [46] and NYU [48] dataset, and hand-object DexYCB [3] datasets. The results show that HandDiff achieves a comparable performance with mean distance errors of 5.76 mm, 6.53 mm, and 7.38 mm on the ICVL, MSRA, and NYU datasets, respectively. The model also significantly outperforms existing state-of-the-art approaches on the DexYCB dataset with the lowest error, achieving the lowest error at 8.06 mm.

The following is a summary of our primary contributions:

- We propose a novel diffusion-based model for hand pose estimation that utilizes the depth image and point cloud input as a multi-modal condition. This model progressively denoises a noise distribution, accurately determining the 3D coordinates of hand joints.
- We propose a novel joint-wise local feature-aware denoising module designed to capture local details surrounding noisy input as a condition for more accurate joint coordinate denoising. Furthermore, this module incorporates a novel kinematic correspondence-aware layer to model the dependencies between joints, thereby enhancing performance.
- We perform comprehensive experiments on big and challenging benchmarks that present the new state-of-the-art performance of our proposed method.

2. Related Work

3D hand pose estimation based on depth image. Among the various 3D hand pose estimation approaches that use depth images, conventional 2D CNN-based methods [6, 11–13, 18, 35, 36, 48] have been widely used due to their simplicity. However, they suffer from limitations such as difficulty in capturing the 3D structure and dependence on the camera's viewpoint.

To overcome these limitations, 3D CNN-based methods [14, 29] were introduced, which use 3D voxelized representations of depth images to capture volumetric information. Although these methods improved the performance of 3D hand pose estimation, they require large amounts of memory and computation, which limits their practical applications.

In contrast, PointNet-based methods [5, 9, 15, 16, 23] process the point cloud, which is an accurate representation of the 3D structure. PointNet [33] is a deep learning framework that can handle irregular and unstructured point clouds. The use of PointNet for hand pose estimation was first introduced in HandPointNet [15]. Point-to-Point model [16] and SHPR-Net [5] further improved performance by generating the point-wise probability distribution. Subsequently, SHPR-Net [5] combined HandPointNet with an auxiliary semantic segmentation subnetwork to enhance performance. Recently, HandFoldingNet [9] introduced a folding concept that reshapes a predefined 2D hand skeleton into hand poses, further improving the estimation accuracy. However, a significant drawback of the point cloud is that querying neighbors from a dense point set for convolution requires heavy computations. Therefore, existing methods commonly use a sparse point cloud, which restricts the performance.

Hence, in this work, we utilize multi-modal representations that combine 2D depth images and 3D point clouds. Thus, the model is able to efficiently extract dense detail information, as well as effectively capture 3D spatial features for accurate 3D hand pose estimation. Moreover, for the first time, we apply the diffusion model with a PointNet-based denoising process to improve pose estimation performance.

Diffusion models for pose estimation. Diffusion models [19, 43], also known as denoising diffusion probabilistic models (DDPMs), are a family of deep generative models. DM recovers the originally observed data distribution from the perturbed data distribution with gradually injected noise by recurrently denoising the noise of each perturbation step. In recent years, they have seen remarkable success in a variety of computer vision tasks, such as object detection [4], image synthesis [20, 39–41], graph generation [22, 30, 51], semantic segmentation [1, 2], and pose estimation [10, 17, 21, 42, 42].

Existing diffusion-based pose estimation approaches [10, 17, 21, 42] have been used mainly for 3D human pose estimation, which regresses the locations of 3D keypoints of humans from 2D RGB images of the human body. This is because the uncertain 2D-to-3D lifting can be modeled as a probability distribution. Specifically, D3DP [42] proposed a multi-hypothesis aggregation with joint-wise reprojection to determine the best hypothesis from the diffusion model using the 2D prior. DiffPoses [17, 21] both introduced a heatmap representation of 2D joints to condition the reverse diffusion process. DiffuPose [10] adopted the graph convolutional network as a denoising function to explicitly learn the connectivity between human joints. In common, these methods all follow the same two-stage scheme, which first requires a trained 2D regression model to obtain 2D keypoints as a prior, and then applies a diffusion model conditioned on the 2D prior to solve the 2D-to-3D lifting problem. Intuitively, the performance of the 2D regression model constrains the denoising quality.

In contrast, our method applies a single-stage denoising process using the conditions from a 3D space. Our method directly accepts raw depth and point cloud frames as conditions in order to take full use of both dense 2D visual features and 3D structural information without requiring any compressed 2D/3D prior information from pre-trained models. Furthermore, our method leverages the DDIM [44] to accelerate inference, since it can complete the thousands of reverse denoising processes in single-digit downsampled timesteps.

3. The Proposed Hand Pose Diffusion Model

HandDiff is a diffusion model that takes a 3D normal distribution and a hand depth image as input and produces the coordinates of the hand joints as output, as shown in Figure 2. Intuitively, HandDiff iteratively removes noise to refine the joint locations by exploring the local region around each joint conditioned on joint-wise features. The input to HandDiff is a hand depth image $\mathbf{D}_{in} \in \mathbb{R}^{H\times W}$ with a set of sampled 3D point coordinates $\mathbf{P}_{in} \in \mathbb{R}^{N\times 3}$, and the outputs are 3D joint coordinates $\mathbf{J}^0 \in \mathbb{R}^{J\times 3}$ denoised from a randomly initialized normal distribution. The depth image and the N points are first supplied into a local condition encoder that extracts local and global features. We construct a ConvNeXt-based autoenoder to generate a 2D local visual feature map $\mathbf{F}_{2d} \in \mathbb{R}^{H/2\times W/2\times d_{2d}}$ and a 2D global vector. Due to the irregularity and disorder of the input point set, we exploit the hierarchical point cloud encoder [25, 34] proposed by PointNet++ [34] to extract 3D local geometric features $\mathbf{F}_{3d} \in \mathbb{R}^{N/2\times d_{3d}}$ and a global 3D vector. Then, the local features are fed into the joint-wise condition extractor to extract joint-wise condition vectors. Afterward, a novel joint-wise local feature-conditioned denoiser is executed T' steps to iteratively denoise joint coordinates by resampling the useful local features around noisy joints and being conditioned by joint-wise conditions.

3.1. Joint-wise Condition Extraction

The joint-wise conditions $\mathbf{C} \in \mathbb{R}^{J\times d_c}$ essentially are the embeddings that represent the joints in the d_c-dimensional latent space. Thus, we take the concept of joint-wise embeddings proposed by HandFoldingNet [9] to obtain the joint-wise conditions. The generation of the joint-wise conditions (joint-wise embeddings) is sequentially performed by a three-layer bias-induced layer (BIL) [8]. The concatenation of the 2D and 3D global vectors is replicated J times and fed into the BILs to generate conditions for the J individual joints. The BIL provides joint-wise independent biases that can be regarded as learnable positional embeddings that enable individual mapping for different joints from the same global feature.

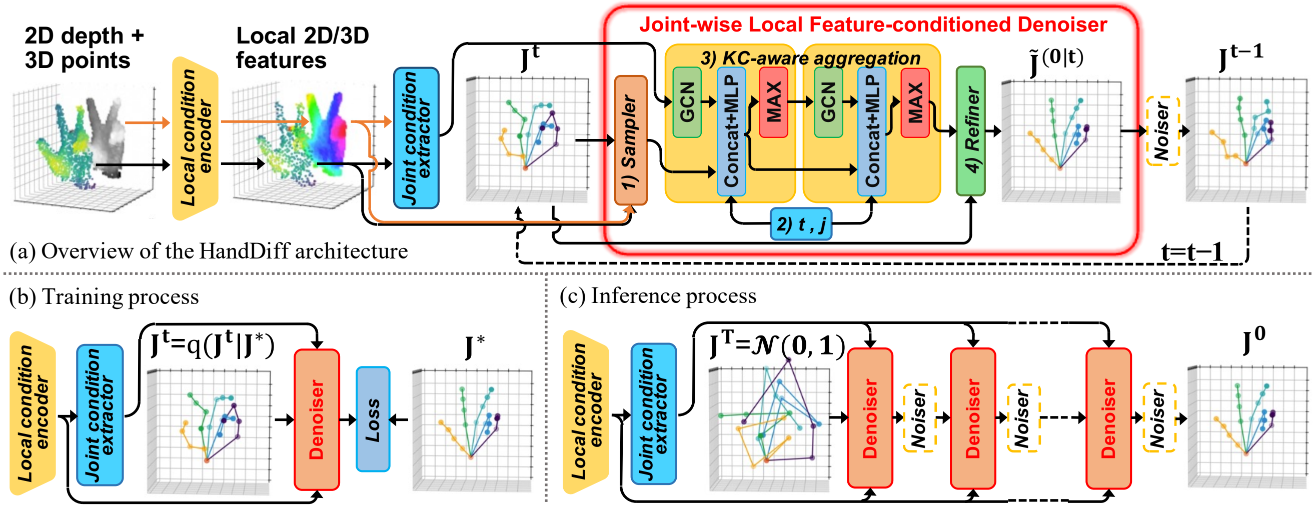
3.2. Joint-wise Local Feature-conditioned Denoiser

The denoiser reconstructs an accurate joint coordinate distribution from a noisy distribution under the joint-wise conditions extracted from the hand point cloud. At each time step t, the denoiser $\mathcal{D}$ accepts the noisy joint coordinate distribution $\mathbf{J}^{(t)}$, joint-wise condition $\mathbf{C}$, local features $\mathbf{F}_{2d}$, $\mathbf{F}_{3d}$ and time step t as input, and outputs the reconstructed 3D joint coordinates $\widetilde{\mathbf{J}}^{(0|t)}$ without noise:

$$\widetilde{\mathbf{J}}^{(0|t)} = \mathcal{D}(\mathbf{J}^{(t)}, \mathbf{C}, \mathbf{F}_{2d}, \mathbf{F}_{3d}, t). \tag{1}$$

The denoiser consists of the following elements: 1) a local feature sampler, 2) a joint indicator & timestep embedding, 3) a kinematic correspondence-aware aggregation block, and 4) a residual refiner. The sampler samples local features around noisy coordinates $\mathbf{J}^{(t)}$ to form local conditions that contain detailed observations. In order to differentiate between different joints and levels of noise, we introduce a joint indicator and a time-step embedding, respectively. Afterward, the aggregation block fuses the local conditions and joint-wise conditions together and subsequently produces the denoised joint coordinates. In addition, the aggregation also cooperates with graph reasoning to capture kinematic correspondences. In the end, a residual mechanism is applied to refine the noisy joint coordinate distribution to the denoised one.

Local feature sampler. The sampler collects K-nearest-neighbor local 3D points and their corresponding local 3D spatial features around each noisy joint distribution. The sampler also samples K-nearest-neighbor 2D pixels and their corresponding local 2D visual features. Note that the 2D pixels are projected onto the same 3D joint space for neighbor querying. The neighbor points and pixels are then



(a) Overview of the HandDiff architecture

(b) Training process

(c) Inference process

Figure 2. The pipeline of the proposed HandDiff. HandDiff takes the normalized point cloud transformed from a 2D depth image as the input. The PointNet-based local condition encoder extracts local features, aka local conditions, from input points. Then, a joint-wise condition extractor aggregates local features into latent features of each joint. Conditioned on the joint-wise conditions and local conditions sampled around each joint, the joint-wise local feature-conditioned denoiser iteratively recovers an accurate 3D hand pose by denoising the diffused noisy pose. Notably, a noiser proposed in DDIM is applied to add noise to the denoised pose for subsequent denoising steps.

translated into a relative coordinate system with the joint location as the origin to eliminate translation variations.

Joint indicator vector & timestep embedding. As the denoiser needs to handle different joints with different levels of noise, it has to be explicitly informed of the joint index j and timestep t. Following DDPMs, we apply the sinusoidal function on j and t to form a joint indicator vector and timestep embedding. Subsequently, the joint indicator vectors and time-step embeddings are concatenated with each local feature for the subsequent process.

Kinematic correspondence-aware aggregation block. As visualized in Figure 2, the proposed aggregation block accepts the local features, joint-wise conditions, joint indicator vectors, and timestep embeddings as input, and outputs kinematic correspondence-evolved local features and joint-wise embeddings. Since many recent approaches [12, 36] suggested that a Graph Convolutional Network (GCN) can effectively model relative kinematic relationships among joints, the joint-wise conditions (embeddings) are first augmented through a GCN, forming the evolved joint-wise conditions:

$$\mathbf{C}' = ReLU(\mathbf{ACW}), \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{d_c \times d_c}$ is the trainable weights and $\mathbf{A} \in \mathbb{R}^{J \times J \times d_c}$ is the channel-wise kinematic correspondence matrix among joints. Afterward, the GCN-evolved joint-wise conditions are concatenated with the joint indicator vectors, timestep embeddings, as well as sampled local features. Subsequently, the concatenated features are sent to a one-layer MLP to generate evolved local features encoding with kinematic correspondence and global prior. Formally, the evolved local feature for the k-th local neighbor point of the j-th joint is defined as

$$\mathbf{F}'_{k,j} = MLP([\mathbf{P}_{k,j} - \mathbf{J}_j, \mathbf{F}_{k,j}, \mathbf{C}'_j, PE(t), PE(j)]), \tag{3}$$

where $\mathbf{P}_{k,j}$ and $\mathbf{F}_{k,j}$ are the coordinate and local feature of the k-th local neighbor point of the j-th joint, PE is the sinusoidal positional embedding function and '$[\cdot,\cdot]$' is the concatenation operation.

However, the single proposed block is not sufficient for complex denoising. Thus, the proposed block is replicated four times with independent learnable parameters in practice. Furthermore, we introduce a max-pooling layer between every two blocks for providing updated joint-wise embeddings with local information to the latter block:

$$\hat{\mathbf{C}} = MaxPool(\mathbf{F}'). \tag{4}$$

Residual refiner. Similar to the previous 3D generative diffusion model [27], the refiner accepts the last joint-wise embeddings $\hat{\mathbf{C}}$ as input to refine the noisy input distribution $\mathbf{J}^{(t)}$. The refiner is a linear transformation with a residual connection. Therefore, the approximated joint coordinates of the current t-th timestep are represented as:

$$\widetilde{\mathbf{J}}^{(0|t)} = \hat{\mathbf{C}}\mathbf{W} + \mathbf{J}^{(t)}, \tag{5}$$

where $\mathbf{W} \in \mathbb{R}^{d_3 \times 3}$ is the trainable transformation matrix.

3.3. Training

HandDiff first corrupts the ground truth joint distribution $q(\mathbf{J}^{(0)})$ to a noisy distribution $q(\mathbf{J}^{(t)}|\mathbf{J}^{(0)})$ by gradually adding noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ through a forward diffusion Markovian chain, where t is uniformly sampled from the predefined total time steps T. Following DDPMs [19], the forward noise process is formally defined as:

$$q(\mathbf{J}^{(t)}|\mathbf{J}^{(0)}) = \sqrt{\bar{\alpha}_t}\mathbf{J}^{(0)} + \epsilon\sqrt{1-\bar{\alpha}_t},$$
$$\text{where } \bar{\alpha}_t = \prod_{s=0}^{t} \alpha_t = \prod_{s=0}^{t}(1-\beta_s). \tag{6}$$

Note that $0 < \beta_t < 1$ is the variance of the noise, which is controlled by a linear variance schedule at each time step, as in DDPM [19].

Subsequently, the noisy joint distribution is supplied to the proposed denoiser to recover the clean joint distribution $\widetilde{\mathbf{J}}^{(0|t)}$, under the joint-wise conditions as well as the local detail conditions. To train the denoiser, $\widetilde{\mathbf{J}}^{(0|t)}$ is under the supervision of the ground truth distribution $\mathbf{J}^*$.

Besides, the joint-wise conditions have to be initialized through training. Therefore, the 3D coordinates $\mathbf{J}^c$ linearly transformed from the joint-wise conditions are also under the same supervision.

Following previous regression works [9, 35], we adopt a smooth L1 loss to supervise training because of its less sensitivity to outliers. The smooth L1 loss is defined as:

$$L1_{smooth}(\mathbf{x}) = \begin{cases} 50\mathbf{x}^2, & |\mathbf{x}| < 0.01 \\ |\mathbf{x}| - 0.005, & otherwise \end{cases}. \tag{7}$$

By using the smooth L1 loss, we supervise the approximated joint distribution by the following joint loss function:

$$\mathcal{L} = \sum_{j=0}^{J} L1_{smooth}(\widetilde{\mathbf{J}}_j^{(0|t)} - \mathbf{J}_j^*). \tag{8}$$

3.4. Inference

During inference, a reverse diffusion process is pursued by iteratively applying the denoiser, to recover the uncontaminated joint coordinate distribution. According to recent 2D-to-3D human pose diffusion models [10, 17, 21, 42], multiple diverse hypotheses for the reverse process can help probabilistic diffusion models to achieve improved accuracy. Our model also samples H initial 3D poses $\mathbf{J}_{0:H}^{(T)}$ from a unit Gaussian distribution.

Afterward, H pose hypotheses are individually passed to the proposed denoiser to approximate the H uncontaminated joint coordinate distribution $\widetilde{\mathbf{J}}_{0:H}^{(0|t)}$. To obtain the noisy input for the subsequent denoising step $t-1$, we exploit a noiser that adds noise to the denoised distribution following the DDIM [44]:

$$p_\theta(\mathbf{J}_{0:H}^{(t-1)}|\widetilde{\mathbf{J}}_{0:H}^{(0|t)}) = \sqrt{\bar{\alpha}_{t-1}}\widetilde{\mathbf{J}}_{0:H}^{(0|t)} + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\epsilon_t + \sigma_t\epsilon, \tag{9}$$

where t is started from T, $\epsilon_t = (\mathbf{J}_{0:H}^{(t)} - \sqrt{\bar{\alpha}_t}\widetilde{\mathbf{J}}_{0:H}^{(0)})/\sqrt{1-\bar{\alpha}_t}$ is the predicted noise of timestep t and $\sigma_t = \sqrt{(1-\bar{\alpha}_{t-1})(1-\bar{\alpha}_t/\bar{\alpha}_{t-1})/(1-\bar{\alpha}_t)}$.

This procedure will be iterated T' times ($T' < T$) to estimate the final denoised distribution $\widetilde{\mathbf{J}}_{0:H}^{(0|t)}$. At the last timestep 0, we average over all hypotheses to aggregate the ultimate uncontaminated joint coordinates:

$$\bar{\mathbf{J}}^{(0)} = \frac{1}{H}\sum_{h=0}^{H}\widetilde{\mathbf{J}}_h^{(0)}. \tag{10}$$

4. Experiments

4.1. Experiment Settings

We conducted experiments on an NVIDIA TITAN RTX GPU with PyTorch. For training, we used the AdamW optimizer [26] with beta$_1$ = 0.5, beta$_2$ = 0.999, and learning rate α = 0.001. The input image was resized to 128, the number of input points to the network was randomly sampled to 1,024, the 2D/3D feature depths d_{2d} and d_{3d} are 128, the joint-wise condition depth d_c is 512 and the batch size was set to 64. The diffusion timestep was set to 500 with a cosine variance scheduler. Meanwhile, to avoid overfitting, we adopted online data augmentation with random rotation ([-180.0, 180.0] degrees), 3D scaling ([0.8, 1.2]), and 3D translation ([-20, 20] mm). We trained the model for 30 epochs with a learning rate decay of 0.1 after every 10 epochs.

4.2. Datasets and Evaluation Metrics

MSRA Dataset. The MSRA dataset [46] provides more than 76K depth image frames, each of which provides $J = 21$ annotated joints, including one joint for the wrist and four joints for each finger. The frames are split into 9 subjects, each of which contains 17 hand gestures.

ICVL Dataset. The ICVL dataset [47] provides 22K training and 1.6K testing depth frames, each of which provides $J = 16$ annotated joints, including one joint for the palm and three joints for each finger.

NYU Dataset. The NYU dataset [48] provides depth images captured from three different views by the PrimeSense 3D sensor. Each view contains 72K frames and 8K frames for training and testing, respectively. Following recent works [9, 15, 16], we use one view for training and testing and selected 14 joints out of a total of 36 annotated joints for evaluation.

DexYCB. The DexYCB dataset [3] is a recently released hand-object dataset that consists of 582,000 image frames with 21 annotated joints, 10 different subjects, and 20 YCB objects from 8 camera views. This dataset defines four official dataset split protocols: S0 - seen subjects, camera views, grasped objects; S1 - unseen subjects; S2 - unseen camera views; S3 - unseen grasped objects.

Evaluation metrics. We employ two commonly used metrics, the mean joint error, and the success rate, to evaluate the performance of hand pose estimation. The mean joint

Table 1. Comparison of the proposed method with previous state-of-the-art methods on the ICVL, MSRA, and NYU datasets. Input indicates the input type of 2D depth image (D), 3D voxels (V), or 3D point cloud (P). † The results are reported from the retrained VVS following the same cropping strategy [31] as in the previous state-of-the-art methods [9, 15, 16, 29, 36, 37, 46].

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

Table 2. Comparison of the proposed method with previous state-of-the-art methods on the DexYCB datasets.

Method	Mean joint error (mm)					Input
	S0	S1	S2	S3	AVG	
A2J [53]	23.93	25.57	27.65	24.92	25.52	D
Spurr et al. [45]	17.34	22.26	25.49	18.44	18.44	RGB
METRO [24]	15.24	-	-	-	-	RGB
Tse et al. [49]	16.05	21.22	27.01	17.93	20.55	RGB
HandOcc [32]	14.04	-	-	-	-	RGB
IPNet [38]	8.03	9.01	8.60	7.80	8.36	D+P
Ours	**7.66**	**8.73**	**8.40**	**7.53**	**8.07**	D+P

error measures the average Euclidean distance between the estimated and ground-truth joint locations for each joint over the testing set. The success rate reveals the percentage of good frames with a mean joint error of less than a certain distance threshold.

4.3. Comparison with State-of-the-Art Methods

Single hand. We compare HandDiff on the ICVL, MSRA, and NYU dataset with other state-of-the-art methods, including methods with 2D depth images as input: improved DeepPrior (DeepPrior++) [31], Pose-Ren [6], dense regression network (DenseReg) [52], CrossInfoNet [11], JGR-P2O [12], spatial-aware stacked regression network (SSRN) [37], pose-guided hierarchical graph network (PHG) [36] and virtual view selection (VVS) [7], and methods with 3D point cloud or voxels as input: 3DCNN [14], SHPR-Net [5], HandPointNet [15], Point-to-Point [16], V2V [29], HandFoldingNet [9] and IPNet [38].

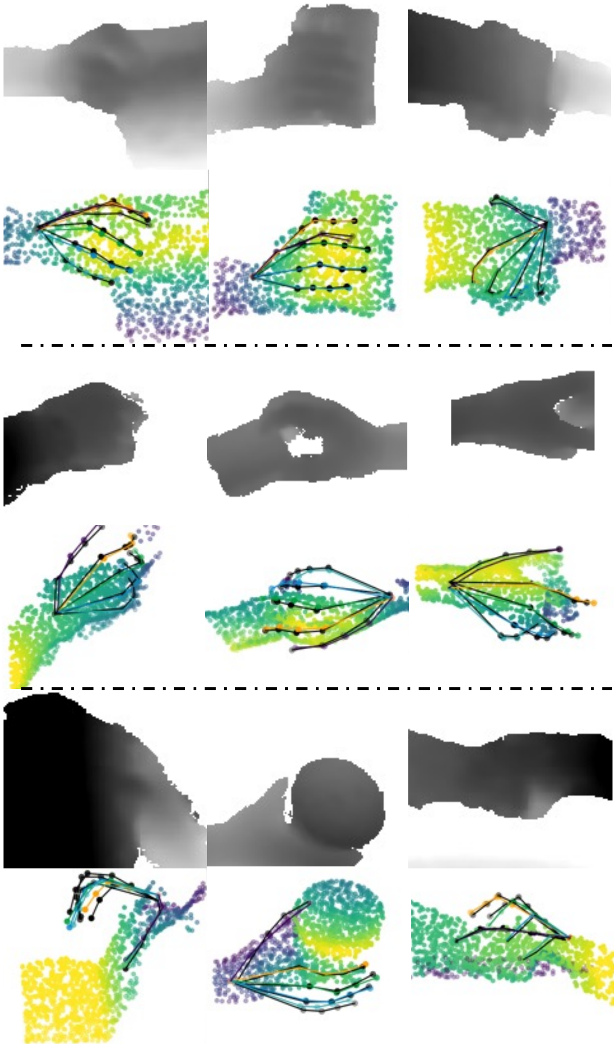


Figure 3. Qualitative results of HandDiff on the DexYCB datasets including different grabbing poses (top), self-occlusions (middle), and object occlusions (bottom). Hand-depth images (first rows) are transformed into 3D points (second rows) in order to clearly present occlusions as shown in the figure. Ground truth is shown in black and the estimated joint coordinates of our model are shown in colors.

Table 1 summarizes the results in terms of the mean joint error on the three datasets. The results show that HandDiff achieves the new state-of-the-art record with mean distance errors of 5.72 and 6.53 mm on two challenging datasets, ICVL and MSRA, respectively. The proposed model also achieves the third-lowest error on the NYU dataset. The results also demonstrate that the proposed HandDiff significantly outperforms other 2D image-based methods by large margins since HandDiff directly performs the processing on the 3D space, avoiding the highly non-linear mapping problem of estimating from the 2D image. Figure 4 illustrates that our method significantly outperforms other methods in terms of success rate when the error threshold is lower than

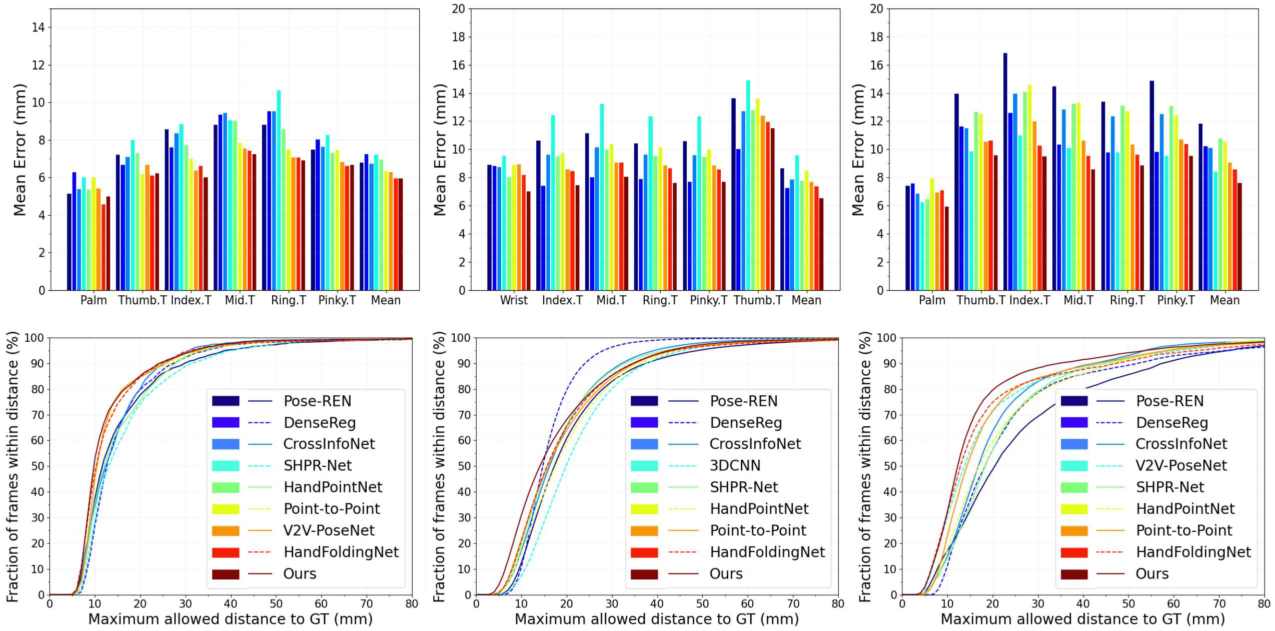


Figure 4. Comparison with the state-of-the-art methods using the ICVL (left), MSRA (middle), and NYU (right) dataset. The per joint error (top) and success rate (bottom) are shown in this figure.

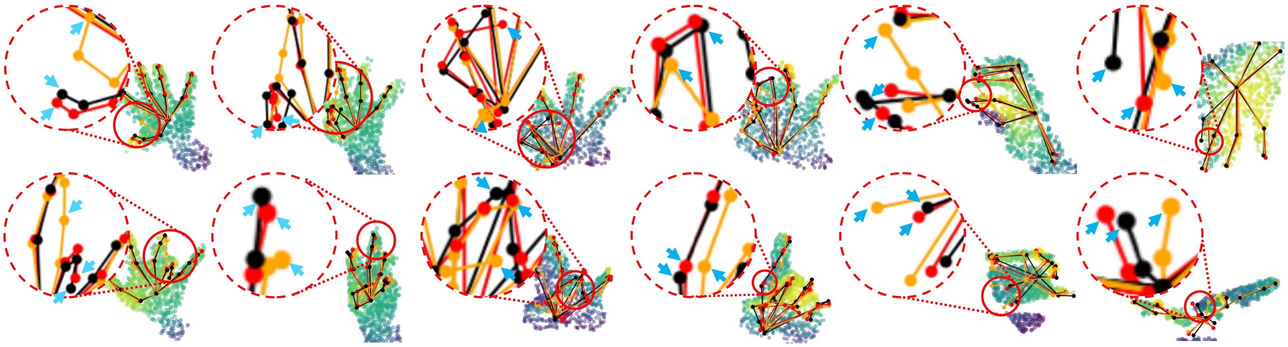


Figure 5. Qualitative results of HandDiff on the ICVL (left), MSRA (middle), and NYU (right) datasets. Hand-depth images are transformed into 3D points in order to clearly present occlusions as shown in the figure. Ground truth is shown in black, results from comparative HandFoldingNet [9] are shown in orange, and the estimated joint coordinates of our model are shown in red.

12, 15, and 52 mm on the ICVL, MSRA, and NYU datasets, respectively.

Hand-object. We compare HandDiff on the hand-object dataset DexYCB with other state-of-the-art method on the official dataset split protocals, including A2J [53], Spurr et al. [45], METRO [24], Tse et al. [49], HandOccNet [32] and IPNet [38]. As shown in Table 2, HandDiff outperforms previous SOTA methods in all four protocols. The qualitative results visualized in Figure 3 also reveal that HandDiff can estimate accurate poses from hand-object interaction scenarios with various occlusions.

4.4. Ablation Study

We conducted extensive ablation experiments to evaluate the contribution of each component proposed in our model.

Analysis of different proposed components. To verify the effectiveness and necessity of the components proposed in this work, we incrementally introduce these components on the existing 3D diffusion probabilistic model (3DDPM) [27], which is able to generate complex point cloud conditionally. Briefly, 3DDPM is a share-weight point-wise denoiser conditioned on a global shape latent. We set the number of its output points as the number of hand joints to adapt it to the hand pose estimation task. Afterward, we follow DDIM for the denoising acceleration. Based on this baseline, we incrementally adopt the proposed components and conduct ablations as follows: 1) using local conditions (LC); 2) using joint indicator (JI); 3) using joint-wise condition (JC) and LC; 4) using LC with JI; 5) using JC and LC with JI; 6) using JC and LC with JI and kinematic cor-

Table 3. Ablations of different proposed components. All the ablation models are trained and tested on the DexYCB dataset .

JC	LC	JI	KC	MH	Mean joint error
		✓			9.17 mm
	✓				49.58 mm
✓					8.37 mm
✓	✓				8.23 mm
	✓	✓			8.28 mm
✓	✓	✓			8.13 mm
	✓	✓	✓		7.94 mm
✓	✓	✓	✓		7.74 mm
✓	✓	✓	✓	✓	**7.66** mm

Table 4. Ablations of different modalities of conditions. All the ablation models are trained and tested on the DexYCB dataset .

Condition modality	Mean joint error (mm)
2D depth	8.23
3D points	9.58
2D depth + 3D points	7.74

respondence (KC); 7) using JC, LC with JI and KC, and multiple hypotheses (MH), which is our full configuration. Note that the use of local conditions without KC is implemented by applying a PointNet layer [33] on noisy joints.

Table 3 reports the experimental results of the ablations. The results demonstrate that the proposed local condition is permutation-equivariant and thus cannot work solely. The joints must be generated in a specific permutation in order to match the permutation defined by the dataset. Therefore, the proposed joint indicator and joint-wise condition that introduce permutation information are mandatory to improve performance. With the help of the joint indicator and joint-wise condition, the proposed local condition mechanism can significantly reduce the mean joint error by more than 0.9 mm and 0.1 mm, respectively. Furthermore, the proposed kinematic correspondence improves performance by learning the inter-joint relations. Finally, the multiple hypotheses further boost the accuracy.

Modality of conditions. As the quality of conditions determines the quality of pose denoising, we feed the diffusion model with different models of input. Table 4 shows that the model combining both 2D and 3D conditions presents the optimal estimation performance. The results also reveal that the model with only 2D conditions slightly degrades because of the 3D information loss. On the other hand, the model with only 3D conditions cannot capture dense features from only 1024 points, thus the estimation error significantly increases.

Number of denoising timesteps. As suggested by DDIM [44], the inference process follows a non-Markovian chain. Thus, the number of denoising timesteps can vary for accelerated inference. Figure 6 (top) visualizes the mean joint errors with the increasing number of timesteps during inference (w/o multiple hypotheses). The results show that the model can approach to an acceptable mean joint error with two-step diffusion. The reason is intuitive that the quantity of hand keypoints to be denoised is relatively small compared to other heavy image/point cloud denoising tasks, which normally require hundreds of timesteps. The results also show that 10 timesteps appear as the optimal 7.44 mm. Larger timesteps exhibit a negligible impact on performance. In addition, the computation time and memory of the model are 98 ms and 2.2GB per frame, respectively, for 10 timesteps (1 hypothese).

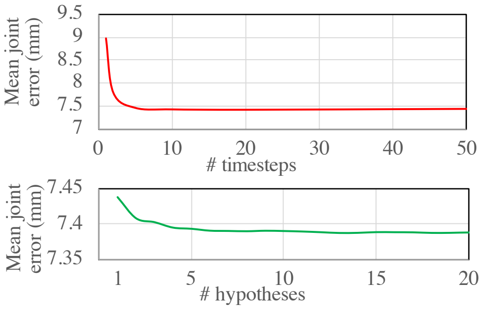


Figure 6. Evaluation results of the increasing number of timesteps (top) and hypotheses (bottom) in the denoising process on the NYU dataset. The model is trained with 500 diffusion timesteps.

Number of hypotheses. Figure 6 (bottom) shows the mean joint errors with the different number of hypotheses for denoising. As expected, the error decreases as the hypothese amount increases. However, the improvement becomes marginal when the number is larger than 10.

5. Conclusion

This paper presented HandDiff, a novel diffusion-based architecture that is capable of reconstructing accurate 3D hand pose iteratively, conditioned on both depth image and point cloud. Experimental results showcased that our network significantly outperforms previous state-of-the-art methods on four challenging datasets. Extensive experiments also reveals the effectiveness of the components proposed in this paper. However, a limitation of HandDiff is its inability to handle scenarios with interacting hands. Future research avenues could explore extensions to bipartite graph learning and skeleton-based analysis to address these limitations and further enhance the model's capabilities.

Acknowledgement. This work was supported by the Technology Innovation Program (RS-2023-00235718) funded by the Ministry of Trade, Industry & Energy (1415187474), and by the Institute of Information and Communication Technology Planning Evaluation (IITP) grants (IITP-2021-0-02068, IITP-2021-0-02052, IITP-2023-2020-0-01821, IITP-2019-0-00421). Wencan Cheng was partly supported by the China Scholarship Council (CSC).

References

[1] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021. 3

[2] Emmanuel Asiedu Brempong, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer, and Mohammad Norouzi. Denoising pretraining for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4175–4186, 2022. 3

[3] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 2, 5

[4] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788*, 2022. 3

[5] Xinghao Chen, Guijin Wang, Cairong Zhang, Tae-Kyun Kim, and Xiangyang Ji. Shpr-net: Deep semantic hand pose regression from point clouds. *IEEE Access*, 6:43425–43439, 2018. 1, 2, 6

[6] Xinghao Chen, Guijin Wang, Hengkai Guo, and Cairong Zhang. Pose guided structured region ensemble network for cascaded hand pose estimation. *Neurocomputing*, 395:138–149, 2020. 1, 2, 6

[7] Jian Cheng, Yanguang Wan, Dexin Zuo, Cuixia Ma, Jian Gu, Ping Tan, Hongan Wang, Xiaoming Deng, and Yinda Zhang. Efficient virtual view selection for 3d hand pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 419–426, 2022. 6

[8] Wencan Cheng and Sukhan Lee. Point auto-encoder and its application to 2d-3d transformation. In *International Symposium on Visual Computing*, pages 66–78. Springer, 2019. 3

[9] Wencan Cheng, Jae Hyun Park, and Jong Hwan Ko. Handfoldingnet: A 3d hand pose estimation network using multiscale-feature guided folding of a 2d hand skeleton. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11260–11269, 2021. 1, 2, 3, 5, 6, 7

[10] Jeongjun Choi, Dongseok Shim, and H Jin Kim. Diffupose: Monocular 3d human pose estimation via denoising diffusion probabilistic model. *arXiv preprint arXiv:2212.02796*, 2022. 3, 5

[11] Kuo Du, Xiangbo Lin, Yi Sun, and Xiaohong Ma. Crossinfonet: Multi-task information sharing based hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9896–9905, 2019. 1, 2, 6

[12] Linpu Fang, Xingyan Liu, Li Liu, Hang Xu, and Wenxiong Kang. Jgr-p2o: Joint graph reasoning based pixel-to-offset prediction network for 3d hand pose estimation from a single depth image. In *European Conference on Computer Vision*, pages 120–137. Springer, 2020. 1, 4, 6

[13] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3593–3601, 2016. 2

[14] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. 3d convolutional neural networks for efficient and robust hand pose estimation from single depth images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1991–2000, 2017. 2, 6

[15] Liuhao Ge, Yujun Cai, Junwu Weng, and Junsong Yuan. Hand pointnet: 3d hand pose estimation using point sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8417–8426, 2018. 1, 2, 5, 6

[16] Liuhao Ge, Zhou Ren, and Junsong Yuan. Point-to-point regression pointnet for 3d hand pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 475–491, 2018. 1, 2, 5, 6

[17] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. *arXiv preprint arXiv:2211.16940*, 2022. 3, 5

[18] Hengkai Guo, Guijin Wang, Xinghao Chen, Cairong Zhang, Fei Qiao, and Huazhong Yang. Region ensemble network: Improving convolutional network for hand pose estimation. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 4512–4516. IEEE, 2017. 1, 2

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 2, 5

[20] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23(47):1–33, 2022. 3

[21] Karl Holmquist and Bastian Wandt. Diffpose: Multi-hypothesis human pose estimation using diffusion models. *arXiv preprint arXiv:2211.16487*, 2022. 3, 5

[22] Jaehyeong Jo, Seul Lee, and Sung Ju Hwang. Score-based generative modeling of graphs via the system of stochastic differential equations. In *International Conference on Machine Learning*, pages 10362–10383. PMLR, 2022. 3

[23] Shile Li and Dongheui Lee. Point-to-pose voting based hand pose estimation using residual permutation equivariant layer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11927–11936, 2019. 1, 2

[24] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1954–1963, 2021. 6, 7

[25] Xingyu Liu, Charles R Qi, and Leonidas J Guibas. Flownet3d: Learning scene flow in 3d point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 529–537, 2019. 3

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[27] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021. 1, 2, 4, 7

[28] Zhaoyang Lyu, Zhifeng Kong, Xudong Xu, Liang Pan, and Dahua Lin. A conditional point diffusion-refinement paradigm for 3d point cloud completion. *arXiv preprint arXiv:2112.03530*, 2021. 1

[29] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pages 5079–5088, 2018. 2, 6

[30] Chenhao Niu, Yang Song, Jiaming Song, Shengjia Zhao, Aditya Grover, and Stefano Ermon. Permutation invariant graph generation via score-based generative modeling. In *International Conference on Artificial Intelligence and Statistics*, pages 4474–4484. PMLR, 2020. 3

[31] Markus Oberweger and Vincent Lepetit. Deepprior++: Improving fast and accurate 3d hand pose estimation. In *Proceedings of the IEEE international conference on computer vision Workshops*, pages 585–594, 2017. 6

[32] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1496–1505, 2022. 6, 7

[33] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 2, 8

[34] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, pages 5099–5108, 2017. 3

[35] Pengfei Ren, Haifeng Sun, Qi Qi, Jingyu Wang, and Weiting Huang. Srn: Stacked regression network for real-time 3d hand pose estimation. In *BMVC*, page 112, 2019. 1, 2, 5

[36] Pengfei Ren, Haifeng Sun, Jiachang Hao, Qi Qi, Jingyu Wang, and Jianxin Liao. Pose-guided hierarchical graph reasoning for 3-d hand pose estimation from a single depth image. *IEEE Transactions on Cybernetics*, 2021. 1, 2, 4, 6

[37] Pengfei Ren, Haifeng Sun, Weiting Huang, Jiachang Hao, Daixuan Cheng, Qi Qi, Jingyu Wang, and Jianxin Liao. Spatial-aware stacked regression network for real-time 3d hand pose estimation. *Neurocomputing*, 437:42–57, 2021. 6

[38] Pengfei Ren, Yuchen Chen, Jiachang Hao, Haifeng Sun, Qi Qi, Jingyu Wang, and Jianxin Liao. Two heads are better than one: image-point cloud network for depth-based 3d hand pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2163–2171, 2023. 1, 6, 7

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3

[40] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.

[41] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[42] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. *arXiv preprint arXiv:2303.11579*, 2023. 3, 5

[43] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 1, 2

[44] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 5, 8

[45] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *European conference on computer vision*, pages 211–228. Springer, 2020. 6, 7

[46] Xiao Sun, Yichen Wei, Shuang Liang, Xiaoou Tang, and Jian Sun. Cascaded hand pose regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 824–832, 2015. 2, 5, 6

[47] Danhang Tang, Hyung Jin Chang, Alykhan Tejani, and Tae-Kyun Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3786–3793, 2014. 2, 5

[48] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (ToG)*, 33(5):1–10, 2014. 2, 5

[49] Tze Ho Elden Tse, Kwang In Kim, Ales Leonardis, and Hyung Jin Chang. Collaborative learning for hand and object reconstruction with attention-guided graph convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1664–1674, 2022. 6, 7

[50] Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, Karsten Kreis, et al. Lion: Latent point diffusion models for 3d shape generation. *Advances in Neural Information Processing Systems*, 35:10021–10039, 2022. 1, 2

[51] Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation. *arXiv preprint arXiv:2209.14734*, 2022. 3

[52] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Dense 3d regression for hand pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5147–5156, 2018. 6

[53] Fu Xiong, Boshen Zhang, Yang Xiao, Zhiguo Cao, Taidong Yu, Joey Tianyi Zhou, and Junsong Yuan. A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 793–802, 2019. 6, 7

[54] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5826–5835, 2021. 1