HHMR: Holistic Hand Mesh Recovery by Enhancing the Multimodal Controllability of Graph Diffusion Models

Mengcheng Li[1], Hongwen Zhang[2], Yuxiang Zhang[1], Ruizhi Shao[1], Tao Yu[1], Yebin Liu[1].
[1]Tsinghua University [2]Beijing Normal University.

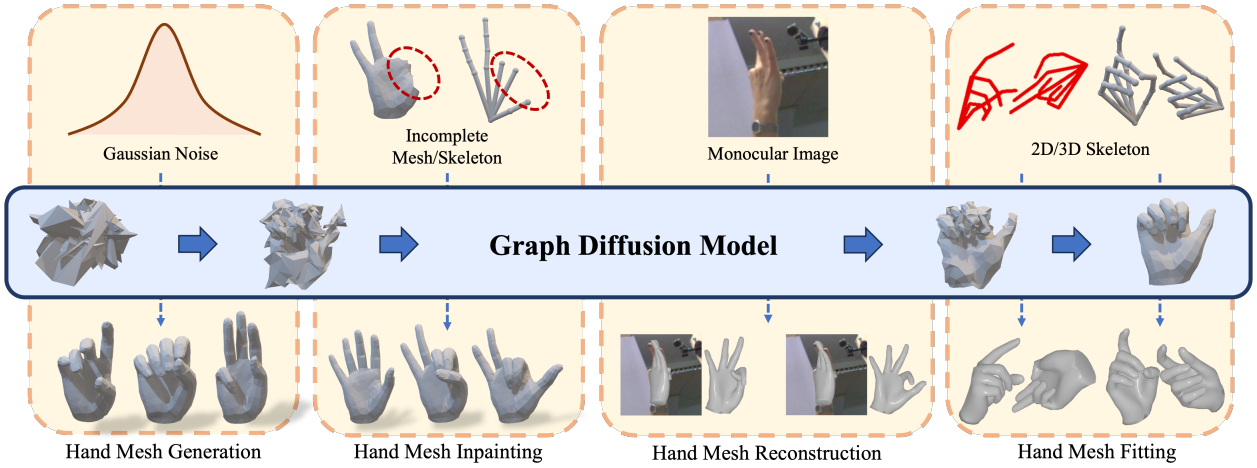

Figure 1. We introduce **HHMR**, a graph diffusion-based generation framework that are compatible with various human hand mesh recovery tasks.

Abstract

*Recent years have witnessed a trend of the deep integration of the generation and reconstruction paradigms. In this paper, we extend the ability of controllable generative models for a more comprehensive hand mesh recovery task: direct hand mesh generation, inpainting, reconstruction, and fitting in a single framework, which we name as **H**olistic **H**and **M**esh **R**ecovery (HHMR). Our key observation is that different kinds of hand mesh recovery tasks can be achieved by a single generative model with strong multimodal controllability, and in such a framework, realizing different tasks only requires giving different signals as conditions. To achieve this goal, we propose an all-in-one diffusion framework based on graph convolution and attention mechanisms for holistic hand mesh recovery. In order to achieve strong control generation capability while ensuring the decoupling of multimodal control signals, we map different modalities to a shared feature space and apply cross-scale random masking in both modality and feature levels. In this way, the correlation between different modalities can be fully exploited during the learning of hand priors. Furthermore, we propose Condition-aligned Gradient Guidance to enhance the alignment of the generated model with the control signals, which significantly improves the accuracy of the hand mesh reconstruction and fitting. Experiments show that our novel framework can realize multiple hand mesh recovery tasks simultaneously and outperform the existing methods in different tasks, which provides more possibilities for subsequent downstream applications including gesture recognition, pose generation, mesh editing, and so on.*

1. Introduction

Hand mesh recovery aims to recover hand pose and mesh from images, which has achieved a wide range of applications in ARVR, human-computer interaction, robotics, and etc,. In the past decades, with the development of deep learning techniques, the paradigm of hand mesh recovery has undergone a shift from keypoints-based mesh&model fitting [14, 25, 46, 50, 57] to learning-based regression. Regression-based hand mesh recovery methods are based

on the encoder-decoder architecture which first encodes various kinds of inputs, usually images or keypoints, and then decodes to either 3D poses [44], pose and shape parameters [2, 6, 55, 58] of a parametric hand model [38] or the vertex coordinates of a 3D hand mesh directly [4, 5, 15, 27, 30, 31, 45]. However, the one-way regression strategy does not establish or fully utilize the underlying distributions of natural hand pose and mesh, which significantly restricts the achievement of more advanced hand mesh recovery tasks, such as mesh&pose inpainting or generation.

Despite the classical hand mesh recovery task, another spectrum of research focuses on formulating the priors (or distributions) of natural hand pose or mesh by low-dimensional manifolds[48]. Acquiring such a prior can not only facilitate the hand mesh recovery but also enable the generation of hands that ideally follow the biomechanical constraints. Classical hand priors either rely on handcrafted 3D hand models with restricted joint angles [44] and degrees of freedom (DoFs) [53]. Romero *et al*. [38] introduced an explicit parametric hand model by principal component analysis (PCA) on a dataset with natural hand pose and shape parameters. However, the limited representative power of explicit constraints or PCA significantly limits the ability to represent complex postures not to say generate diverse hand meshes.

Benefiting from the rapid development of generative learning techniques, generation-based implicit hand models become a hot topic. Some works [28, 61] have utilized the Variational Autoencoder (VAE) to map the plausible hand parameters to a Gaussian space. For the first time, it becomes possible to generate hand meshes through sampling in feasible distributions. Implicit hand models can also be used for hand mesh recovery. By using two VAEs to encode the image domain and hand pose domain independently, CrossVAE or Generative Adversarial Networks (GAN) are incorporated to align the two domains and enable the generation of hand parameters that align with the input image ([43, 49, 51, 52]). However, existing implicit hand models are mostly built on the parameter space or 3D skeleton space, which makes it hard to generate meshes or utilize geometric constraints directly. Moreover, additional inversion or fitting processes are necessary for hand mesh recovery based on existing implicit hand models, which is time-consuming and sophisticated.

An inevitable trend for related research is the unification of hand mesh recovery and the generative hand models. This will enhance the mutual benefits between existing reconstruction and generation paradigms and finally enable direct hand mesh generation, inpainting, reconstruction, and fitting simultaneously in an End2End manner. We call such a comprehensive hand mesh recovery task as **H**olistic **H**and **M**esh **R**ecovery, named HHMR, and propose an All-In-One diffusion framework based on GCN and attention for holistic hand mesh recovery without any additional finetuning or inversion steps. Our key insight is that conditional generation based on diffusion models, through carefully multimodal designed and conditioned training, is a powerful methodology for achieving holistic hand mesh recovery. Given different modalities of conditions as input, the diffusion model can produce plausible hand mesh and pose results thus naturally supports: i) hand mesh reconstruction when given image as condition, ii) hand mesh inpainting when given incomplete mesh or 2D&3D skeleton as condition, iii) hand mesh generation when given random noise as input (with no condition), and iv) hand mesh fitting when given a 2D skeleton as the condition. To fulfill the accurate multimodal controllability of the diffusion models, we map different conditions to a shared feature space and apply a random mask strategy at both modality and feature levels to enhance the correlation learning between different modalities. Moreover, we propose a condition-aligned gradient guidance strategy during diffusion learning to further improve the alignment with the conditions.

We evaluate the proposed holistic hand mesh recovery framework on various downstream tasks. Empirical results demonstrate that i) on hand mesh generation task, our approach generates more diverse poses. ii) On hand mesh inpainting task, our method can recover multiple plausible hand meshes from incomplete inputs. iii) Our method achieved comparable results with SOTAs on single-hypothesis reconstruction and outperforms SOTAs on multi-hypothesis reconstruction tasks. iv) With condition-aligned gradient guidance, our method achieves performance with higher precision in 2D fitting tasks.

The contributions can be summarized as:

- We propose an All-In-One framework based on a graph diffusion model for holistic hand mesh recovery. The hand prior learned in the graph diffusion model can be readily applied to different downstream tasks without any additional finetuning or inversion steps.
- We map different modality conditions to a shared feature space and apply a random mask strategy at both modality and feature levels to enhance the correlation learning between different modalities.
- We propose a condition-aligned gradient guidance strategy during inference to further improve the alignment with the conditions.
- Extensive experiments and comparisons validate the effectiveness of our framework and demonstrate various downstream applications.

2. Related Work

2.1. Hand Mesh Recovery

In the past few decades, significant progress has been made in human hand recovery. Early research [14, 25, 46, 50, 57] utilized optimization methods to fit the hand pose from 2D skeletons detection.

With the development of neural networks, some learning-based hand mesh reconstruction methods have been proposed. One approach is utilizing the popular parameter-based MANO [38] model and regressing the pose and shape parameters of it. Due to the differentiability of the MANO model, it becomes feasible to estimate model parameters end-to-end from a single-view image input([2, 6, 55, 58]). However, parameter estimation is a highly nonlinear task that lacks the correlation with the 3D space. Another approach is directly regressing hand mesh, which is typically achieved by building a mapping network between 2D image input and 3D hand mesh output. Graph convolution network (GCN) is one of the widely employed mapping networks ([4, 5, 15, 27, 45]). Additionally, there are methods that utilize one-dimensional lixel heatmaps [33] to represent the 3D coordinates of vertices, or employ UV map [3] to establish connections between 2D and 3D. Lin *et al*. [30, 31] used transformer-based network to regress vertices.

However, these hand reconstruction methods only provide a single plausible estimation, whereas in reality, the occluded parts could assume various poses. Thus, we propose a probability diffusion-based model that generates multiple plausible hand meshes from one input RGB image.

2.2. Hand Prior and Generation

The hand prior model is aimed at learning a distribution of plausible hand poses. Previous prior models can be broadly categorized into two types: one type learns the unconditional distribution $p_{data}(x)$ of the human hand, and the other is the task-specific prior that learning the conditional distribution $p_{data}(x|c)$ under given conditions (such as RGB image or 2D skeleton).

For the unconditional prior distribution, one straightforward approach is to manually constrain the reasonable range of hand poses based on the biological structure of the human hand. Yang *et al*. [53] have constructed the hand pose prior by manually defining the degrees of freedom for each joint and the range of rotation angles. Spurr *et al*. [44] proposed a set of losses that constrain hand pose to lie within the range of biomechanically feasible 3D hand configurations. The other line of works is learning-based which involves learning feasible distribution from a large dataset of poses. Javier *et al*. [38] applied principal component analysis (PCA) on the pose and shape parameters of a hand dataset. In both human hand domain [28, 61] and body domain [35], researches have been conducted to map the pose distribution to a standard Gaussian distribution by an VAE network. Tiwari *et al*. [48] described the prior of the human body by learning the neural distance from the sampled pose to the low-dimensional manifold of reasonable pose distributions. These unconditional priors often require optimization when applied in downstream applications, which can be time-consuming.

The conditional prior model often learns the potential pose distribution under specific task constraints, such as predicting plausible pose from RGB images or 2D skeletal. A common approach is to construct VAEs in different domains (*e.g*. depth image, RGB image or 3D skeleton) and then build hand priors through latent space alignment([43, 49, 51, 52]). Recently, Ci *et al*. [7] predicted the gradient of log-likelihood of the human body pose distribution through a score matching based model. However, these implicit conditional prior models are mostly built on the 3D skeletons and rely on specific conditions, which restricts their application.

2.3. Diffusion based generative model

The diffusion model [41] is a generative model based on stochastic diffusion processes, which uses a Markov chain to gradually convert one distribution into another as modeled in Thermodynamics. In practice, it pre-defines a forward process that gradually adds noise to a dataset distribution until it converges to a standard Gaussian noise distribution. Subsequently, a neural network is trained to learn the reverse process and gradually denoise a random noise into a reasonable data sample. Recently, some image generation algorithms [23, 37, 42] that utilize diffusion models have achieved significant breakthroughs. Diffusion models have also demonstrated great success in the fields of 3D object generation [29, 36, 39], human pose estimation [13, 16, 24], and human motion generation [40, 47]. We build a diffusion-based hand generation model with a classifier-free strategy, which can run with or without conditional. Hence, it is capable of handling various downstream tasks.

3. Preliminary: Diffusion Model

The denoising diffusion model is a probabilistic generative model that consists of a forward process and a reverse process. We denote the middle distribution for each step in the process as $p(x_t), t = 0, 1, ..., T$, where the initial $p(x_0)$ is the dataset distribution we want to generate, and the final $p(x_T) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a stand Gaussian noise distribution.

The forward process converts $p(x_0)$ to $p(x_T)$ by gradually adding small Gaussian noises with predefined means and variances in a Markov Chain:

$$p(x_t|x_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

where $\{\beta_t\}$ is a predefined set of small constants.

The reverse process is a generation process that converts the noise sampled from $p(x_T)$ to a reasonable data sample in $p(x_0)$. The reverse process is also a Gaussian distribution [12]. Moreover, the reverse diffusion process can be written as:

$$p_\theta(x_{t-1}|x_t, \mathbf{c}) = \mathcal{N}(\mu_t^\theta(x_t, t, \mathbf{c}), \Sigma_t \mathbf{I}) \tag{2}$$

where $\mathbf{c}$ represents the conditions for generation and μ_t^θ, Σ_t are means and variances of the reverse Gaussian process. Usually, Σ_t is determined by $\{\beta_t\}$, while μ_t is untrackable. Thus, a neural network $\mu_t^\theta(x_t, t, \mathbf{c})$ is applied to predict the means μ_t.

Furthermore, for a more effective learning of μ_t^θ, it is typically reparameterized to learn i) the noise ϵ added from x_0 to x_t or ii) x_0 instead. Image generation tasks [10, 22, 23, 42] usually employ the first setting, whereas we follow the human motion generation methods [40, 47] to predict x_0 directly:

$$\bar{x}_0 = f_\theta(x_t, \mathbf{c}, t), \quad \mu_t^\theta(x_t, t, \mathbf{c}) = g_t(\bar{x}_0, x_t, t), \tag{3}$$

where f is a neural network, θ represents the training parameters, g_t is a determined reparameterized function. More details can be referred to [23].

In summary, the training algorithm for the diffusion model involves four parts: i) sample data x_0 from dataset $p(x_0)$, ii) run the forward process to add noise on x_0 and yield x_t, iii) run the neural network f_θ to give a prediction of $\bar{x_0}$, iv) back propagate the prediction loss $\|\bar{x}_0 - x_0\|$.

4. Method

4.1. Problem Statement

Our goal is to develop a diffusion model that learns a prior knowledge of the human hand from the given task-specific conditions $f_\theta(x_t, \mathbf{c}, t)$. Considering the highly non-linear mapping from the parameter space of the hand template to 3D spatial space, we directly process on the 3D meshes instead. We drew inspiration from the network architecture of image generation method [37], but employed 3D graph convolutions network (GCN) [8] instead of 2D image convolutions to encode the local information of the human hand mesh. The details of the network architecture will be presented in the Sec.4.2.

In order to accommodate different downstream tasks, our approach supports a variety of conditions $\mathbf{c}$ as input. Specifically, we set $\mathbf{c}$ to a RGB image $\mathbf{c}_{image}$ to handle monocular hand mesh reconstruction task; set $\mathbf{c}$ to 2D skeleton $\mathbf{c}_{skel2d}$ for 2D hand mesh fitting task; set $\mathbf{c}$ to incomplete 3D skeleton $\mathbf{c}_{skel3d}$ for hand mesh inpainting task. We can also set $c = \emptyset$ to achieve unconditional generation, and apply it to the task of hand postures generation and hand mesh completion.

Due to the different input conditions, our network can be trained using different types of datasets. For the dataset containing paired human hand images and annotated hand meshes, we can train our network under all the aforementioned conditions: $\mathbf{c} = \mathbf{c}_{image}|\mathbf{c}_{skel2d}|\mathbf{c}_{skel3d}$. For the dataset containing only human hand poses, we employ skeletal information $\mathbf{c} = \mathbf{c}_{skel2d}|\mathbf{c}_{skel3d}$ as the condition to train our generation model.

4.2. Graph Diffusion Network

Network architecture. As shown in Fig. 2, the main structure of our diffusion model is a U-shaped network, consisting of a downsampling encoder and an upsampling decoder with skip connections. Each block of both encoder and decoder are formed by 4 layers: *i.e.*, GCN layer, self-attention layer, cross-attention layer, and optional upsampling or downsampling layer.

Similar to the convolution in 2D image networks, we use GCN [8] to aggregate information from each vertex's neighbors to encode local information of hand mesh, and expand the receptive field through stacking multiple layers of GCN. Then a self-attention layer is utilized to facilitate long-range global information propagation on the 3D hand mesh, enabling our prior network to construct global relationships across the surface vertices. We also apply a cross-attention layer to inject conditional information $\mathbf{c}$ into our generative model.

Condition Mapping. In order to encode different conditions into the same space that are compatible with the input of the cross-attention layer, we designed distinct networks to map different conditions to a shared feature space:

- For 3D skeleton and 2D skeleton condition, we simply use MLP to encode $\mathbf{c}_{skel2d}$ and $\mathbf{c}_{skel3d}$.
- For RGB image condition $\mathbf{c}_{image}$, we first use a CNN-based network to encode the latent image feature map and then crop it into several image patches. After that, a Vision Transformer (ViT) [11] based model is applied to further process image information. We also add a global token $\mathbf{c}_{global}$ while running ViT to encode the global information as the original ViT does. The global token is further stacked with the output image patch features $\mathbf{c}_{patch}$ of ViT to form the image condition $\mathbf{c}_{image} \in \mathcal{R}^{(P+1)\times D}$, where P is the number of image patches.

Conditions with Random Mask. In our method, we stack all three types of condition $\mathbf{c} = \mathbf{c}_{image} \oplus \mathbf{c}_{skel2d} \oplus \mathbf{c}_{skel3d} \in \mathcal{R}^{(P+1+1+1)\times D}$ as the input of cross-attention layers. To enhance the network's generalization and versatility, we design a two-level masking strategy. On the multimodal level, in order to decouple the relationship between conditions, we independently set each condition $\mathbf{c}_{image}|\mathbf{c}_{skel2d}|\mathbf{c}_{skel3d}$ to an empty set $\emptyset$ with probability

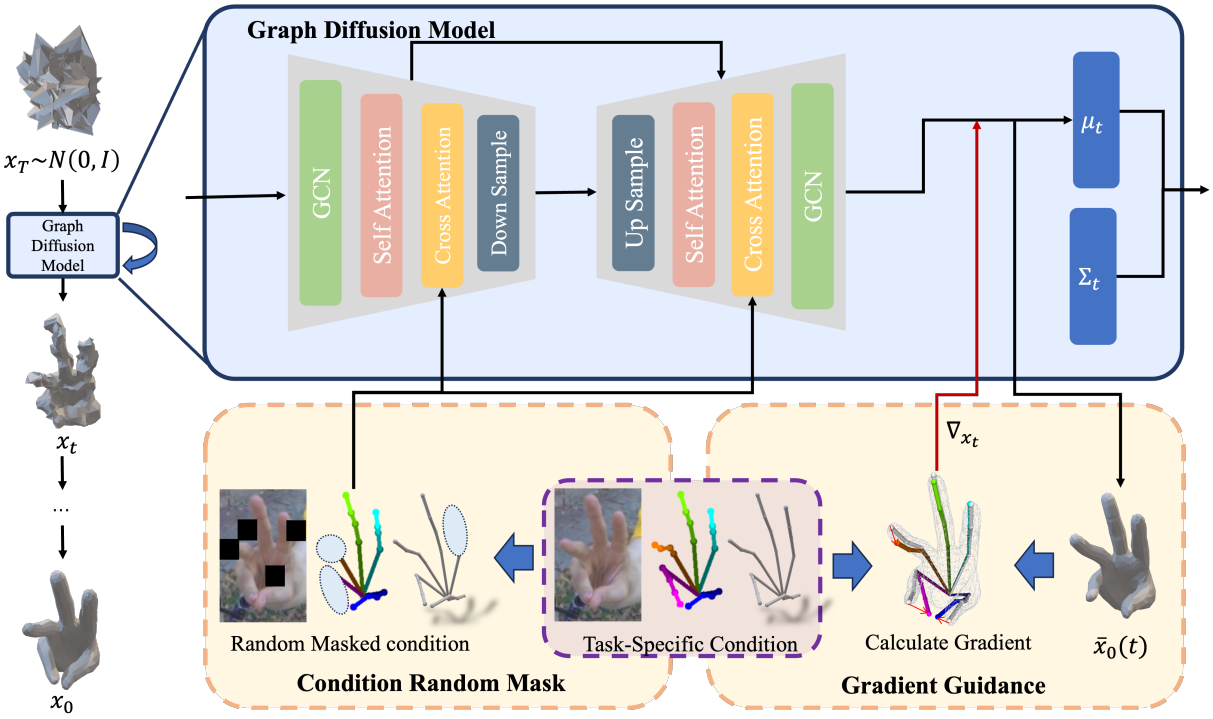

Figure 2. The pipeline of our graph diffusion model. With task-specific conditions, our model progressively removes noise from randomly Gaussian noise and directly reconstructs the complete hand meshes. Additionally, we introduce a gradient-based guidance to improve the alignment between the generated results and observations.

p_m during training. Furthermore, we also globally mask all conditions with probability p_{all} to train our model's unconditional generation capability. On the feature level, We design different masking strategies for each condition individually. For image condition $\mathbf{c}_{image}$, we encode it into several image patches and randomly mask each patch with probability p_{image}, which enables our method to handle complex occlusion input. For skeleton condition $\mathbf{c}_{skel2d}$ and $\mathbf{c}_{skel3d}$, we add Gaussian noise on each joint and randomly mask each finger with probability p_{skel}, which facilitates the generation from incomplete hand skeletons.

Note that when applying our model to a specific task, the irrelevant conditions can be directly set to $\emptyset$. For example, in the monocular reconstruction task, the condition is $\mathbf{c} = \mathbf{c}_{image} \oplus \emptyset \oplus \emptyset$, which we write as $\mathbf{c} = \mathbf{c}_{image}$ for short.

4.3. Condition-aligned Gradient Guidance

In the 2D image generation task, a classifier guidance [10] approach is proposed for the conditional diffusion model. It trains an additional classifier network to predict the probability of input condition from x_t. Then, by modifying the means of the reverse process with a gradient of the log-likelihood predicted from the classifier network, the diffusion model can generate images that are more consistent with the input conditions.

Inspired by previous work [10], we propose a Condition-aligned Gradient Guidance to encourage that the generated hand is consistent with the input condition. Specifically, in our implementation, we add a gradient guidance bias to the means of each Gaussian distribution of the reverse process:

$$\bar{\mu}_t = \mu_t - s\Sigma_t \nabla_{x_t} \|P f_\theta(x_t, \mathbf{c}, t) - P x_0\| \quad (4)$$

where s is a scale factor and P is a task-specific operator. Noted that $f_\theta(x_t, \mathbf{c}, t)$ is the prediction of the GT mesh x_0, this method can be seen as a form of neural fitting. The operator P determines the specific fitting target. For example, by setting P as a joint regression matrix $\mathcal{J}$, the gradient guidance works as a supervise term to constrain the 3D joints of generated hand meshes to be consistent with the 3D skeleton condition.

In addition to serving as a supervisory, this gradient-based guidance approach can also be viewed as a geometric control during the generation process. For example, in the hand mesh inpainting task, given an incomplete hand mesh, we can apply the gradient guidance on the given part and leave the missing part uncontrolled. In this situation, the operator P is a per-vertex binary mask M_V that indicates the missing part of hand meshes. We can also use partial skeletons to control the generation of the human hand by setting P as $M_J\mathcal{J}$, where M_J is a per-joint binary mask.

Note that in contrast to the original classifier guidance method, our gradient guidance approach does not need any extra classifier network. Hence, our approach requires no additional training, which is less time-consuming and plug-to-use.

4.4. Loss Functions

For training our graph diffusion model, we utilize (1) diffusion data loss, (2) vertex loss & joint loss and (3) mesh smooth loss.

Data Loss. We use L1 loss to supervise the output of our diffusion model as described in Sec.3:

$$L_{data} = \|x_0 - \bar{x}_0\| \tag{5}$$

Vertex Loss & Joint Loss. We pretrained a joint regression matrix $\mathcal{J}$ to regress joints from output hand mesh and apply L1 loss to supervise 3D coordinates of vertices and joints:

$$\begin{aligned} L_V &= \sum \|V - V^{GT}\|_1 \\ L_J &= \sum \|\mathcal{J}V - \mathcal{J}V^{GT}\|_1 \end{aligned} \tag{6}$$

Mesh Smooth Loss. To ensure the geometric smoothness of the predicted vertices, two different smooth losses are applied. First, we regularize the normal consistency between the predicted and the ground truth mesh:

$$L_n = \sum \|e \cdot n^{GT}\|_1, \tag{7}$$

where e is the edges of the generated hand mesh and n^{GT} is the faces normal vector calculated from the ground truth mesh. Also, we minimize the L1 distance of each edge length between the generated mesh and the ground truth mesh:

$$L_e = \sum \|e - e^{GT}\|_1. \tag{8}$$

5. Experiments

5.1. Experimental Setting

Implementation Details. Our network is implemented by PyTorch. For RGB image condition, we use ResNet50 [19] with initial weights pretrained on ImageNet [9] as the backbone for the mapping network and then evenly crop the image feature into 8×8 patches. For 2D/3D skeleton conditions, we use a three-layer MLP with dropout [21] and GELU [20] activation. For the diffusion model, we set the denoising steps to 1000 during training, and utilized the DDIM [42] algorithm to speed up inference.

Training Datasets. We simultaneously leveraged hand pose datasets with only hand pose annotations and reconstruction datasets with both RGB images and hand pose annotations. For pose datasets, we use Two-hand 500K [61] and InterHand2.6M [34] datasets. Two-hand 500K [61] is a two-hand pose dataset that utlized hand instances sampled from single-hand datasets [34, 54, 59, 60]. InterHand2.6M [34] contains both single-hand and two hands motion sequence data in 30FPS. we split the pose data of two hands into two individual single-hand data. For reconstruction datasets, we use FreiHand [60], HO3D_V3 [17, 18], CompHand [4, 5], both of them contain hand images and hand mesh annotations.

Training Details. We train our model with a mixture of the above datasets. We train our model using the AdamW optimizer [32] on a single NVIDIA RTX 4090 GPU with a mini-batch size 64. We trained for 1M mini-batches in total. The learning rate is setting to $5e^{-4}$ and decrease to $1e^{-5}$ with cosine annealing.

5.2. Hand Mesh Generation

Our model can work without any condition by setting $\mathbf{c} = \emptyset$. In this setting, our approach involves sampling a set of noisy point clouds from a Gaussian distribution and denoising them into a complete human hand mesh as shown in Fig.3. To quantitatively evaluate the diversity of our generative model, we report the average pairwise distance (APD) [1], which is the mean vertex distance between all pairs of samples. We randomly sample $n = 500$ hand meshes with our model and the result APD is $16.17mm$. For comparison, we also sample $n = 500$ hand meshes within the PCA space of hand parameters from MANO [38] and the result APD is $13.42mm$. The result of APD shows that our model generates more diverse poses than PCA.

5.3. Hand Mesh Inpainting

Given an incomplete hand mesh or hand 3D skeleton, our model can inpaint the whole plausible hand mesh according to the prior knowledge learned from a large hand pose dataset. The key idea of mesh inpainting is using the gradient guidance to keep the given part of the hand unchanged while allowing the diffusion model to give various generations for the missing part. As illustrated in the Fig.3, where the thumb, index, and middle fingers are all missing from the hand mesh, our model can provide various potential completion results. Also, given a 3D hand skeleton that missing the middle, ring, and little fingers, our model can generate diverse whole-hand meshes. Please refer to the Supplementary for the details of the mesh inpainting algorithm.

5.3.1 Hand Mesh Reconstruction

Single-Hypothesis. Given an RGB image as condition $\mathbf{c}_{image}$, our model can work as a monocular hand mesh reconstruction network. We first quantitatively evaluate our method on Freihand [60] dataset. We use DDIM [42] algorithm to run $T = 10$ denoising steps and only sample one

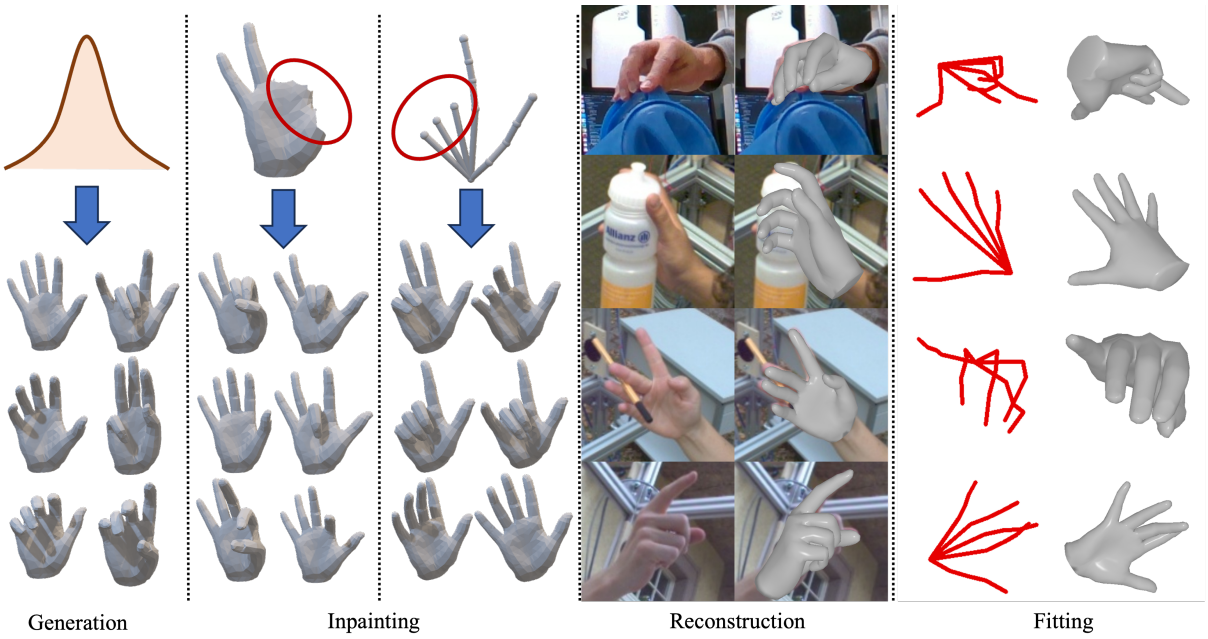

Figure 3. Qualitative results of our method for different downstream tasks. From left to right are i) hand mesh generation results from random Gaussian noise, ii) hand mesh inpainting from incomplete hand mesh or skeleton, iii) hand mesh reconstruction from monocular RGB image, and iv) hand mesh fitting from 2D skeletons.

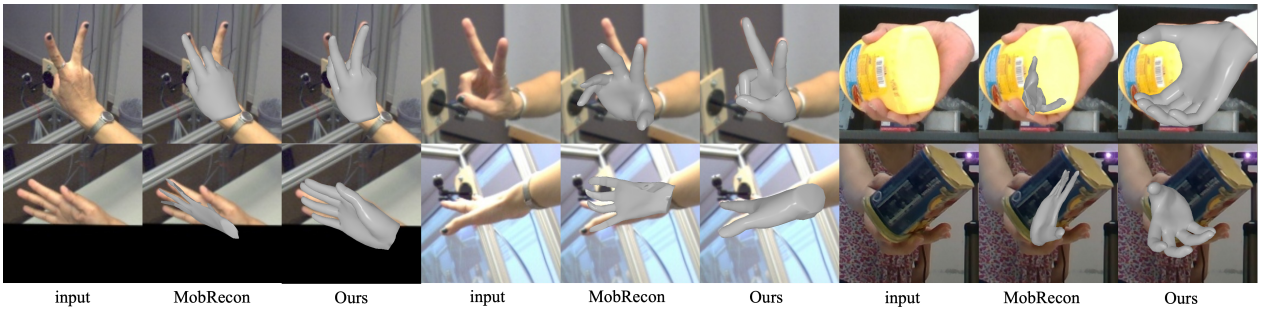

Figure 4. Qualitative results comparison with MobRecon [5]. Left 4 cases are taken from evaluation set of Freihand [60]. Right 2 cases are taken from HO3D_V3 [17, 18].

hypothesis. Note that we follow MeshGraphormer [30] to use the test time augmentation during evaluation. We report the Mean Per Joint Position Error (MPJPE) and the Mean Per Vertex Position Error (MPVPE) after rigid alignment in Tab.1, it can be seen that our method can achieve comparable results with state-of-the-art (SOTA) methods for single-hypothesis reconstruction (n=1). We also conducted qualitative comparisons on the Freihand dataset and HO3D_V3 [17, 18] dataset. As shown in Fig.4, our method performs better on side viewpoint situations and occlusions situations. We believe that this is due to the fact that our model is a prior model, which can better recover hand meshes from incomplete observations.

Multi-Hypothesis. By denoising different noises sampled from Gaussian distribution, our method is capable of providing multiple mesh reconstruction hypotheses from a single image. We first qualitatively demonstrate the multi-hypothesis results in Fig.5, showing that our method can maintain alignment with the input image in visible areas while offering multiple potential guesses for the occluded regions. Then, we quantitatively analyze the effect of the hypothesis count in Tab.2. We also evaluate the effect of inference steps on DDIM [42]. Following human body multi-hypothesis methods [7, 26], we report the minimum MPJPE and MPVPE. As shown in Tab.2 and Tab.1, when only 8 samples are drawn, our method already outperforms the SOTAs [5, 30]. Additionally, Tab.2 also indicates that satisfactory results can be achieved with just $T = 10$ steps

	MPJPE	MPVPE
FreiHAND [60]	11.0	10.9
YotubeHand [27]	8.4	8.6
I2L-MeshNet [33]	7.4	7.6
HIU-DMTL [56]	7.1	7.3
CMR [4]	16.9	7.0
I2UV-HandNet [3]	6.7	6.9
METRO [31]	6.7	6.8
Tang *et al*. [45]	6.7	6.7
MobRecon [5]	6.1	6.2
MeshGraphormer [30]	5.9	6.0
Ours (n=1)	6.0	6.0
Ours (n=8)	5.9	5.9
Ours (n=16)	**5.8**	**5.8**

Table 1. Quantitative results on the Freihand dataset. Our method applied $T = 10$ steps denoising by DDIM [42] and n represents the number of samples.

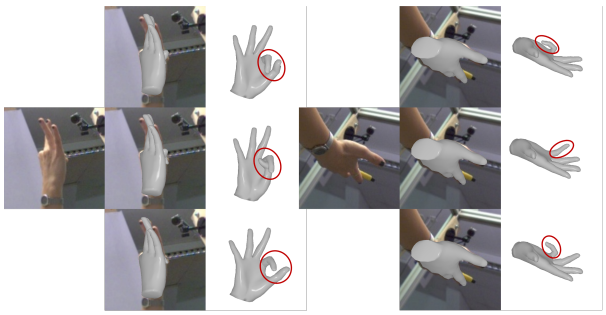

Figure 5. Results for multi-hypothesis reconstruction from monocular image.

		MPJPE	MPVPE
T=10	n=8	5.91	5.90
	n=16	5.76	5.76
	n=32	5.63	5.64
T=25	n=8	5.79	5.79
	n=16	5.58	5.59
	n=32	5.41	5.43
T=50	n=8	5.73	5.73
	n=16	5.51	5.52
	n=32	**5.33**	**5.35**

Table 2. Effect of denoising steps and sampling quantity. T represents the number of denoising steps by DDIM. n represents the number of samples.

of denoising.

5.3.2 Hand Mesh Fitting

Our method can also take a 2D skeleton as a condition $\mathbf{c}_{skel2d}$, and can also utilize the gradient guidance approach

	MPVPE	MPJPE
Fitting 2D skeleton	7.00	7.27
$\mathbf{c}_{skel2D} + \nabla_{skel2D}$	6.28	6.26
$\mathbf{c}_{image}$	6.54	6.57
$\mathbf{c}_{image} + \mathbf{c}_{skel2D}$	5.32	5.29
$\mathbf{c}_{image} + \mathbf{c}_{skel2D} + \nabla_{skel2D}$	**5.05**	**4.94**

Table 3. Quantitative results for hand mesh fitting, $\mathbf{c}_{image}$ represents input RGB image, $\mathbf{c}_{skel2D}$ represents input 2D skeleton, ∇_{skel2D} represents applying the gradient guidance on 2D skeleton

to further enhance alignment. We set the operator P in Equ.4 as $P = M\Pi\mathcal{J}$, where $\mathcal{J}$ is a pre-trained skeleton regression matrix, Π is the projection operator, M is a per-joint mask that masks out 2D joints with low detection confidence. We employ the ground truth 2D skeleton as supervision for quantitative evaluation under various inputs.

We firstly compare with the traditional 2D skeleton fitting algorithm by utilizing the hand PCA prior [38]. We utilize 2D skeleton $\mathbf{c}_{skel2D}$ as input conditional and also apply the gradient guidance ∇_{skel2D}. As demonstrated by the second and third rows in Tab.3, our method achieved higher fitting accuracy. This is because our method is a 3D prior model that is capable of learning the relationship from 2D skeleton to 3D hand mesh.

We also introduce the 2D skeleton information $\mathbf{c}_{skel2D}$ into the hand mesh reconstruction task with image input $\mathbf{c}_{image}$ to further improve the accuracy by 2D fitting. Quantitative results in Tab.3 demonstrate that our fitting method can further improve the alignment with the input image compared to the reconstruction results. please refer to the Supplementary Materials for more qualitative results.

6. Discussion

Conclusion. We introduce a graph diffusion based generative network that is compatible with various downstream hand mesh recovery tasks, including hand mesh generation, hand mesh inpainting, hand mesh reconstruction and hand mesh fitting. Our network can take different task-specific conditions as input, and directly denoise a 3D hand mesh from randomly sampled noisy point clouds. We also designed a conditional mask training strategy to address missing and noisy conditional inputs, and further employed a gradient guidance method for enhancing the consistency between the generation output and the input conditions.

Limitation. Although our method uses a mask training strategy to work with noisy and incomplete conditional inputs, it may still fail for extremely missing or excessively noisy input conditions. Additionally, while our method only requires $T = 10$ denoising steps to achieve decent results, increasing the denoising steps for more precise generation is time-consuming.

References

[1] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[2] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10843–10852, 2019. 2, 3

[3] Ping Chen, Yujin Chen, Dong Yang, Fangyin Wu, Qin Li, Qingpei Xia, and Yong Tan. I2uv-handnet: Image-to-uv prediction network for accurate and high-fidelity 3d hand mesh modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12929–12938, 2021. 3, 8

[4] Xingyu Chen, Yufeng Liu, Chongyang Ma, Jianlong Chang, Huayan Wang, Tian Chen, Xiaoyan Guo, Pengfei Wan, and Wen Zheng. Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration. In *CVPR*, 2021. 2, 3, 6, 8

[5] Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20544–20554, 2022. 2, 3, 6, 7, 8

[6] Yujin Chen, Zhigang Tu, Di Kang, Linchao Bao, Ying Zhang, Xuefei Zhe, Ruizhi Chen, and Junsong Yuan. Model-based 3d hand reconstruction via self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10451–10460, 2021. 2, 3

[7] Hai Ci, Mingdong Wu, Wentao Zhu, Xiaoxuan Ma, Hao Dong, Fangwei Zhong, and Yizhou Wang. Gfpose: Learning 3d human pose prior with gradient fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4800–4810, 2023. 3, 7

[8] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *NeurIPS*, 2016. 4

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 4, 5

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 4

[12] William Feller. Retracted chapter: On the theory of stochastic processes, with particular reference to applications. In *Selected Papers I*, pages 769–798. Springer, 2015. 4

[13] Runyang Feng, Yixing Gao, Tze Ho Elden Tse, Xueqing Ma, and Hyung Jin Chang. Diffpose: Spatiotemporal diffusion model for video-based human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14861–14872, 2023. 3

[14] Chengying Gao, Yujia Yang, and Wensheng Li. 3d interacting hand pose and shape estimation from a single rgb image. *Neurocomputing*, 474:25–36, 2022. 1, 3

[15] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3D hand shape and pose estimation from a single RGB image. In *CVPR*, 2019. 2, 3

[16] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13041–13051, 2023. 3

[17] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020. 6, 7

[18] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *CVPR*, 2022. 6, 7

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[20] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 6

[21] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012. 6

[22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 4

[24] Karl Holmquist and Bastian Wandt. Diffpose: Multi-hypothesis human pose estimation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15977–15987, 2023. 3

[25] Ludovic Hoyet, Kenneth Ryall, Rachel McDonnell, and Carol O'Sullivan. Sleight of hand: perception of finger motion from reduced marker sets. In *Proceedings of the ACM SIGGRAPH symposium on interactive 3D graphics and games*, pages 79–86, 2012. 1, 3

[26] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11605–11614, 2021. 7

[27] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4990–5000, 2020. 2, 3, 8

[28] Jiaman Li, Ruben Villegas, Duygu Ceylan, Jimei Yang, Zhengfei Kuang, Hao Li, and Yajie Zhao. Task-generic hierarchical human motion prior using vaes. In *2021 International Conference on 3D Vision (3DV)*, pages 771–781. IEEE, 2021. 2, 3

[29] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 3

[30] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, 2021. 2, 3, 7, 8

[31] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 2, 3, 8

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[33] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 752–768. Springer, 2020. 3, 8

[34] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3D interacting hand pose estimation from a single rgb image. In *ECCV*, 2020. 6

[35] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 3

[36] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2022. 3

[37] 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, 4

[38] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. In *SIGGRAPH Asia*, 2017. 2, 3, 6, 8

[39] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 3

[40] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 3, 4

[41] 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. 3

[42] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 3, 4, 6, 7, 8

[43] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *CVPR*, 2018. 2, 3

[44] 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. 2, 3

[45] Xiao Tang, Tianyu Wang, and Chi-Wing Fu. Towards accurate alignment in real-time 3D hand-mesh reconstruction. In *ICCV*, 2021. 2, 3, 8

[46] Jonathan Taylor, Vladimir Tankovich, Danhang Tang, Cem Keskin, David Kim, Philip Davidson, Adarsh Kowdle, and Shahram Izadi. Articulated distance fields for ultra-fast tracking of hands interacting. *TOG*, 2017. 1, 3

[47] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 3, 4

[48] Garvita Tiwari, Dimitrije Antić, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision*, pages 572–589. Springer, 2022. 2, 3

[49] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Crossing nets: Combining gans and vaes with a shared latent space for hand pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 680–689, 2017. 2, 3

[50] Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Rgb2hands: real-time tracking of 3d hand interactions from monocular rgb video. *ACM Transactions on Graphics (ToG)*, 39(6):1–16, 2020. 1, 3

[51] Linlin Yang and Angela Yao. Disentangling latent hands for image synthesis and pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9877–9886, 2019. 2, 3

[52] Linlin Yang, Shile Li, Dongheui Lee, and Angela Yao. Aligning latent spaces for 3d hand pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2335–2343, 2019. 2, 3

[53] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11097–11106, 2021. 2, 3

[54] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. A hand pose tracking benchmark from stereo matching. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 982–986. IEEE, 2017. 6

[55] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2, 3

[56] Xiong Zhang, Hongsheng Huang, Jianchao Tan, Hongmin Xu, Cheng Yang, Guozhu Peng, Lei Wang, and Ji Liu. Hand image understanding via deep multi-task learning. In *ICCV*, 2021. 8

[57] Wenping Zhao, Jinxiang Chai, and Ying-Qing Xu. Combining marker-based mocap and rgb-d camera for acquiring high-fidelity hand motion data. In *Proceedings of the ACM SIGGRAPH/eurographics symposium on computer animation*, pages 33–42, 2012. 1, 3

[58] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5346–5355, 2020. 2, 3

[59] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, pages 4903–4911, 2017. 6

[60] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max J. Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *ICCV*, 2019. 6, 7, 8

[61] Binghui Zuo, Zimeng Zhao, Wenqian Sun, Wei Xie, Zhou Xue, and Yangang Wang. Reconstructing interacting hands with interaction prior from monocular images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9054–9064, 2023. 2, 3, 6