HandBooster: Boosting 3D Hand-Mesh Reconstruction by Conditional Synthesis and Sampling of Hand-Object Interactions

Hao Xu[1] Haipeng Li[2] Yinqiao Wang[1] Shuaicheng Liu[2] Chi-Wing Fu[1]
[1]The Chinese University of Hong Kong
[2]University of Electronic Science and Technology of China
{xuhao,yqwang,cwfu}@cse.cuhk.edu.hk, {lihaipeng@std.,liushuaicheng@}uestc.edu.cn

Abstract

Reconstructing 3D hand mesh robustly from a single image is very challenging, due to the lack of diversity in existing real-world datasets. While data synthesis helps relieve the issue, the syn-to-real gap still hinders its usage. In this work, we present HandBooster, a new approach to uplift the data diversity and boost the 3D hand-mesh reconstruction performance by training a conditional generative space on hand-object interactions and purposely sampling the space to synthesize effective data samples. First, we construct versatile content-aware conditions to guide a diffusion model to produce realistic images with diverse hand appearances, poses, views, and backgrounds; favorably, accurate 3D annotations are obtained for free. Then, we design a novel condition creator based on our similarity-aware distribution sampling strategies to deliberately find novel and realistic interaction poses that are distinctive from the training set. Equipped with our method, several baselines can be significantly improved beyond the SOTA on the HO3D and DexYCB benchmarks. Our code will be released on https://github.com/hxwork/HandBooster_Pytorch.

1. Introduction

The task of reconstructing 3D hand mesh from a single image facilitates a wide range of applications, *e.g.*, in AR/VR and human-computer interactions. Recently-proposed data-driven methods show promising results in hand-object interaction scenarios. Yet, their performance is largely limited by the training data, since existing datasets typically lack diversity in hand appearances/poses, views, *etc*.

Existing real-world hand-object datasets are collected in laboratory or in-the-wild scenes. For laboratory-captured datasets such as DexYCB [8] and HO3D [22], they offer a large quantity of hand-object interaction samples with relatively accurate 3D annotations. However, the variations in the samples are still limited, since the data collection process is typically very tedious with the expensive MoCap system. Conversely, for in-the-wild datasets, *e.g.*, YouTube-Hands [40], the data has richer variations, but they provide only pseudo labels, without accuracy guarantee.

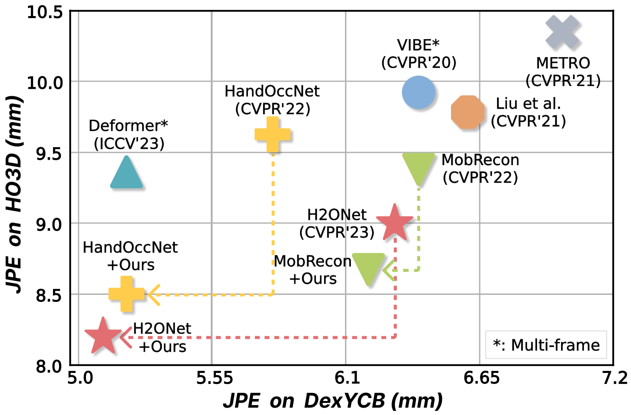

Figure 1. HandBooster significantly improves several baselines, making them SOTA again on both HO3D and DexYCB.

Synthetic data, both rendering- and generative-based, is a workaround to avoid tedious data collection while ensuring annotation precision. ObMan [23], YCB-Afford [17], and ArtiBoost [71] use Blender [16], Maya [1], or PyRender [47] to produce hand renderings. However, the results are unrealistic due to simulated hand appearance and inconsistency in foreground/background lighting. Also, they need extensive extra data such as HDR environment maps for lighting and background augmentation. On the other hand, generative methods, *e.g.*, HOGAN [31], can alleviate these issues. Yet, they consider only novel view synthesis but not other aspects of data diversity. Also, there is no evidence that the hand-mesh reconstruction performance can be improved consistently on existing methods. To sum up, the key challenge is how to generate *realistic and diverse hand-object interaction images with reliable annotations*.

In this paper, we present HandBooster to uplift the data diversity and boost 3D hand-mesh reconstruction *by train-*

[1] Department of Computer Science and Engineering; Institute of Medical Intelligence and XR (IMIXR).

ing a conditional generative space on hand-object interactions and *purposely sampling the space to produce effective data samples*. In short, we first construct versatile content-aware conditions to guide a diffusion model to produce realistic image samples. By this means, we can controllably produce images of diverse hand appearances, poses, views, and backgrounds, where precise 3D annotations are available for free. Further, we design a novel condition creator based on our similarity-aware distribution sampling strategies to find novel interaction poses distinct from the existing ones, thus maximizing the training data quality.

First, directly mapping an input 3D mesh to 2D RGB values is challenging. So, we decompose this process into two steps: (i) project the 3D mesh into a more interpretable 2D image form while preserving its geometric information; and (ii) utilize the 2D results as conditions on the diffusion model to enable controllable generation of realistic images. Among the candidates for the 2D conditions, inspired by [31], we empirically choose the comprehensible and informative normal map and object texture. Further, observing that small changes in 3D hand orientation can hardly be seen in 2D images, we embed 3D hand orientation as an additional condition to ensure orientation-aware generation.

Second, to enhance the reconstruction performance and generalizability, solely utilizing existing hand-object grasping poses is inadequate. The synthesized hand-object interaction samples should be: (i) realistic, ensuring natural grasping poses; (ii) diverse, encompassing various grasping poses; and (iii) novel, encouraging unseen poses. To achieve (i), we employ an optimization-based method [63] to simulate grasping poses, followed by our validation strategy to find natural poses. To promote (ii) and (iii), we further propose similarity-aware sampling strategies, including an intra-distribution farthest pose sampling strategy to avoid repeated or similar poses within both the real and synthetic data and a cross-distribution sampling strategy to encourage the likelihood of sampling novel-generated poses. Last, we train several baselines using our generated samples together with real-world data, showing that their performance can be significantly improved over the SOTA; see Fig. 1.

Our main contributions are summarized as follows,

- We propose **HandBooster**, a new generative framework that utilizes content-aware conditions to synthesize realistic hand-object images with diverse hand appearances, poses, views, and backgrounds, as well as accurate annotations, to boost the reconstruction performance.
- We design the novel condition creator to effectively produce realistic and diverse novel views and grasping poses, by designing the intra-distribution farthest pose sampling strategy and the cross-distribution sampling strategy.
- Extensive evaluations on two widely-used datasets show that HandBooster achieves consistent performance gain on several methods, setting new SOTA performance.

2. Related Works

3D Hand-Mesh Reconstruction. Extensive research has been conducted on 3D hand-mesh reconstruction from RGB images. Most of them tackle the problem by regressing MANO coefficients [2, 3, 5, 7, 13, 23, 34, 44, 60, 69, 74, 76–79, 81]. Others mainly regress voxels [32, 50, 51, 70], implicit functions [48], and vertices [11, 12, 21, 40, 42, 43]. Despite careful network designs, reconstructing accurate 3D hand meshes from monocular images remains challenging, particularly under severe occlusions. Hence, some recent works [9, 20, 65, 66, 68] attempt to leverage multi-frame information. A very recent work [72] uses prior knowledge in a diffusion model to render object geometries to improve everyday object reconstruction. Our approach is orthogonal to these methods and achieves top performance when partnered with several recent baselines [12, 52, 66].

Hand-Object Interaction Image Synthesis. So far, few works have explored image synthesis of hand-object interactions. These works are either rendering- or generative-based. The former employs rendering tools such as Blender [16] and Maya [1] to produce synthetic data. To generate grasping poses, some [17, 33, 71] design specific algorithms, while others [23] use off-the-shell tools such as GraspIt [49] and recent works [62, 63, 67]. However, when applied to 3D hand-mesh reconstruction, these synthetic data inevitably introduce noticeable domain gaps compared to real-world data, leading to inferior performance.

Generative-based methods can produce more realistic images. HOGAN [31] synthesizes novel views using the target posture as guidance; yet, it cannot effectively generate images of novel grasping poses and lacks other aspects of diversity. Another work [73] generates hand graspings on RGB images that contain objects by using a diffusion model conditioned on a generated hand orientation mask. However, its diversities are still restricted by the input images and the lack of annotations severely limits its usage in downstream applications. In this work, we are able to synthesize realistic and diverse hand-object interactions that encompass various appearances, grasping poses, object types, and camera views, significantly enhancing the hand-mesh reconstruction performance, as shown in Fig. 1.

Conditional Diffusion Model. Diffusion model is a type of generative model that adopts the stochastic diffusion process in thermodynamics [55]. Besides, it can also be formulated as a score-based generative model [57, 58]. Recently, DDPM [28] models complex data distributions through discrete steps. In this work, we focus mainly on conditioned generation, including classifier-guided [18, 45] and classifier-free [27] methods. Though LDM [53], ControlNet [75], or other related methods [4, 29, 30, 35, 37, 41, 46, 54, 56, 64] appear as potential tools for our work, they are primarily designed for producing high-resolution images or

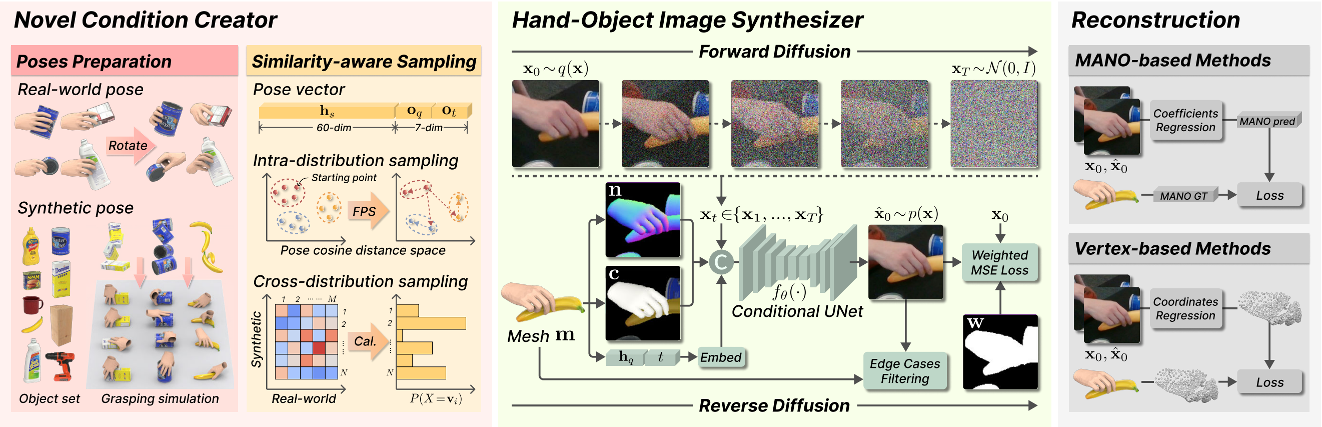

Figure 2. Our HandBooster framework. (i) The Novel Condition Creator prepares and samples diverse and novel grasping poses against real-world and synthetic distributions to create conditions. (ii) The Hand-Object Image Synthesizer follows the conditions to generate image samples. (iii) The synthesized samples can then be employed to effectively train different types of reconstruction models.

exploiting data priors in pre-trained models, lacking task-specific designs to meet the needs of 3D hand-mesh reconstruction, *e.g.*, plausible hand-object interactions and hand orientation awareness. In this work, we propose various content-aware conditions to guide the DDPM, demonstrating controllability and fine-grained realistic results.

3. Method

3.1. Overview

Fig. 2 shows the overall HandBooster framework. First, we formulate the hand-object image synthesizer, which utilizes content-aware conditions to generate compatible hand-object images (Sec. 3.2). Next, to maximize its efficacy, we design the novel condition creator, aiming to prepare and sample realistic, diverse, and novel grasping poses as conditions (Sec. 3.3). Combining their strengths, we can effectively generate realistic hand-object images with annotations to train various reconstruction models (Sec. 3.4.)

3.2. Hand-Object Image Synthesizer

We first describe the process of generating realistic images from a given 3D hand mesh $\mathbf{m}$ using our hand-object image synthesizer. To tackle the challenge of constructing a mapping between the 3D coordinate space and the 2D image space, we decompose the process into two steps, as inspired by [27, 53]. First, we perform a 3D-to-2D projection while retaining sufficient information to construct content-aware conditions. Then, we adopt them to guide a conditional diffusion model to generate realistic hand-object images.

Content-aware Conditions. To minimize the information lost during the 3D-to-2D projection process and reduce the learning difficulty, we aim to choose informative and interpretable conditions for synthesizing hand-object images. We have considered several candidates in 2D format, including skeleton, segmentation, texture, depth, and normal maps, which can be divided into two categories: depth and normal maps contain more topology information, while others are more semantic. To strike a balance, a sound solution is to choose one from each category. As Fig. 3 shows, the texture map contains both shape and color knowledge, so it is selected due to its high informativeness. Also, the normal map is selected for its ease of interpretation, as different parts of the hand and object are more distinguishable compared to the depth map. However, relying solely on 2D images cannot capture small changes in the 3D hand orientation, leading to large regression errors in the camera space during the reconstruction. Hence, we design another condition to encourage orientation-aware generation, where the hand orientation is represented using quaternion, embedded into latent space, and incorporated into several stages of the diffusion model. This inclusion allows for better handling of changes in hand orientation and gives hints to generate arms (not included in conditions) in the image synthesis.

Controllable Realistic Image Synthesis. We first utilize the classifier-free diffusion model [27, 28] for a controllable generation of hand-object images. The forward diffusion is a Markov chain of diffusion steps, in which Gaussian noise is added progressively to a real-world data sample, $\mathbf{x}_0 \sim q(\mathbf{x})$, producing a noisy transition sequence $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T$. To facilitate the training, we use the parameterization method following [38] to sample $\mathbf{x}_t$ at arbitrary timestamp. Then, we adopt a UNet-like classifier-free denoising model $f_\theta(\cdot)$ to learn the reverse diffusion, taking our content-aware conditions to control the generation of the hand-object images from an isotropic Gaussian noise. After the training, we can obtain a controllable model for synthesizing high-quality images with fine-grained details, following our content-aware conditions.

During training, $\mathbf{x}_0 \in \mathbb{R}^{128 \times 128 \times 3}$ are the same image used for 3D hand-mesh reconstruction. To improve the

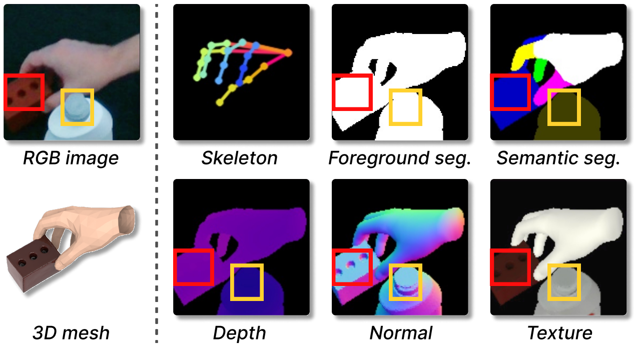

Figure 3. Condition candidates. Normal and texture maps are more informative than others; see the red/yellow bounding boxes.

quality of the generated data and employ them in our task, we incorporate our content-aware conditions $\mathbf{y}$ with $\mathbf{x}_t$ and t to generate samples $\hat{\mathbf{x}}_0$ using the denoising model $f_\theta(\cdot)$,

$$\hat{\mathbf{x}}_0 = f_\theta(\mathrm{Concat}(\mathbf{n}, \mathbf{c}), \mathbf{h}_q, \mathrm{PE}(t)), \tag{1}$$

where $\mathrm{Concat}(\cdot)$ denotes the channel-wise concatenate operation; $\mathbf{n}$ and $\mathbf{c}$ indicate the normal and texture maps, respectively; and $\mathbf{h}_q$ is the hand orientation, represented using quaternion to ensure continuity; and the timestamp t is encoded through positional embedding (PE) [61]. Following [37], we let the denoising model $f_\theta(\cdot)$ predict RGB values and calculate L2 loss with $\mathbf{x}_0$. Formally, we have

$$\mathcal{L}_{DM} = \mathbf{w}\frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}||\hat{\mathbf{x}}_0 - \mathbf{x}_0||_2^2, \tag{2}$$

where $\mathbf{w}$ is a pixel-wise weighting map to enhance the foreground, since the generation quality of the background is less important. We use the hand-object segmentation map as the mask and set the weight to 0.1 for the background.

In the reverse diffusion, synthetic samples $\hat{\mathbf{x}}_0$ are generated from randomly-sampled Gaussian noise $\mathbf{x}_t \in \mathcal{N}(\mathbf{0}, \mathbf{I})$ with the content-aware conditions $\mathbf{n}$, $\mathbf{c}$, and $\mathbf{h_q}$. After certain timestamps, the generated images $\hat{\mathbf{x}}_0$ would become realistic and the input 3D meshes $\mathbf{m}$ are exactly the corresponding ground truths, forming a training pair $(\hat{\mathbf{x}}_0, \mathbf{m})$.

Edge Cases Filtering. Though we can effectively generate realistic images, it is unlikely to fully avoid undesired artifacts, so filtering is necessary. To mitigate the negative effects of training the reconstruction models with such image samples, we propose a selection process to filter out edge cases. Specifically, we adopt a 3D hand-mesh reconstruction model pre-trained on the same training set to evaluate the generated images $\hat{\mathbf{x}}_0$. Here, we calculate the 3D joint and vertex errors between the predicted results and the corresponding ground truths $\mathbf{m}$. Considering that the model should yield reasonable outcomes on eligible synthetic images that align with the distribution of real-world data, we thus exclude edge-case samples that exhibit significant errors over predetermined thresholds in the selection.

3.3. Novel Condition Creator

To effectively generalize the reconstruction models to handle unseen scenarios while avoiding potential over-fitting, it is essential to incorporate diverse hand-object interaction poses. However, simply augmenting existing conditions can only produce novel views. Doing so is insufficient to boost the reconstruction performance. Hence, we design the novel condition creator to construct novel and diverse conditions by finding unseen poses and combining them with the existing ones via our distribution sampling strategies.

Pose Preparation. We first try to fully utilize the grasping poses in the current training set by augmenting their hand orientations. Since not all frames involve object grasping, the data is split into "grasping" and "non-grasping" parts. Only "grasping" poses are augmented to synthesize novel views, while others remain unchanged to preserve realistic motions. Yet, existing datasets [8, 22] do not provide labels for grasping status, so we derive the grasping status from the object pose automatically. We calculate the isotropic Relative Rotation Error (RRE) and Relative Translation Error (RTE) between the initial and current poses:

$$\mathrm{RRE}{=}\arccos(\frac{\mathrm{trace}(\mathbf{o}_r^{t\,T}\mathbf{o}_r^0{-}1)}{2}),\ \mathrm{RTE}{=}||\mathbf{o}_t^t - \mathbf{o}_t^0||_2, \tag{3}$$

where $\mathbf{o}_r$ and $\mathbf{o}_t$ denote the object rotation matrix and translation vector, respectively; and the superscripts 0 and t denote the first and t-th frames, respectively, in the same sequence. If the difference exceeds a pre-defined threshold (empirically setting 5° on RRE and $10mm$ on RTE), the frame is labeled as "grasping". To produce novel views, the orientations of the "grasping" poses are perturbed in a certain range and then rendered as conditions.

Next, we focus on generating novel hand-object grasping poses. Specifically, we use the same YCB [6] objects as in DexYCB [8] and obtain initial poses by simulating the process of falling from a random height to a plane. To generate a grasping pose, we employ the recent work [63] for its fast convergence and high success rate. Please refer to [67] for the details. We perform this simulation 10,000 times on each object. Though penalties have been applied to encourage realism, undesired poses still exist. Thus, we design a validation process to select high-quality poses. A qualified pose must meet three criteria: (i) the hand must make contact with the object (it may fail to grasp small/thin objects); (ii) no obvious hand-object intersection (the volume of the intersection part should be smaller than a pre-defined threshold); and (iii) no self-penetration for the hand. Finally, to avoid introducing domain gaps, we align the orientation of the generated pose to a random one from the training set and apply perturbation as augmentation.

Intra-distribution Sampling. The grasping pose partly depends on the initial object pose. Yet, the number of fea-

Algorithm 1: Farthest Pose Sampling Algorithm

```
Data: Set of poses $\mathbf{P}$, number of sampled poses $M$
Result: Sampled set of poses $\mathbf{Q}$
$\mathbf{Q} \leftarrow \varnothing \cup \{\mathbf{v}_0 \in_R \mathbf{P}\}$;
while $|\mathbf{Q}| < M$ do
    $\mathbf{v}_{max} \leftarrow$ None;
    $d_{max} \leftarrow -\infty$;
    for $\mathbf{v}_i \in \mathbf{P},\ \mathbf{v}_i \notin \mathbf{Q}$ do
        $d_i \leftarrow \min_{\mathbf{v}_j \in \mathbf{Q}} D_c(\mathbf{v}_i, \mathbf{v}_j)$;
        if $d_i > d_{max}$ then
            $d_{max} \leftarrow d_i$;
            $\mathbf{v}_{max} \leftarrow \mathbf{v}_i$;
        end
    end
    $\mathbf{Q} \leftarrow \mathbf{Q} \cup \{\mathbf{v}_{max}\}$;
end
```

sible grasping poses is inherently limited by the shape of the object, *e.g.*, a bowl can only be placed upright or upside down on a plane, each poses having very limited ways of grasping. Though it is possible to grasp the same object multiple times, the resulting poses could be too similar to one another, as different grasping poses do not occur with equal probability. To promote pose diversity and avoid cases that dominate the distribution, inspired by [19], we propose a farthest-pose-sampling (FPS) strategy to select poses as evenly as possible for each object category.

For real-world and synthetic grasping pose sets $\mathbf{P}_r$ and $\mathbf{P}_s$, we perform this intra-distribution sampling separately to obtain the sampled sets $\mathbf{Q}_r$ and $\mathbf{Q}_s$. The subscript is ignored for brevity. The pose vector $\mathbf{v}$ is constructed as

$$\mathbf{v} = \text{Concat}(\mathbf{h}_s, \mathbf{o}_q, \mathbf{o}_t), \quad \mathbf{v} \in \mathbb{R}^n \tag{4}$$

where $\mathbf{h}_s$ and $\mathbf{o}_q$ denote the quaternion representation of the hand pose and object rotation, respectively. Note that the global orientation is removed from the grasping pose, which means we compute the distance at the canonical pose. As described in Alg. 1, FPS begins by initializing the sampled pose set $\mathbf{Q}$ using a random sample $\mathbf{v}_0$ from the input pose set $\mathbf{P}$. While the number of samples in $\mathbf{Q}$ is less than M (M and N for real-world and synthetic data, respectively), it traverses the remaining poses in $\mathbf{P}$ and selects $\mathbf{v}_{max}$ with the largest nearest distance d_{max} to $\mathbf{Q}$ iteratively. The distance between two pose vectors $\mathbf{v}_i \in \mathbf{P}, \mathbf{v}_i \notin \mathbf{Q}$ and $\mathbf{v}_j \in \mathbf{Q}$ is computed using the cosine distance $D_c(\cdot,\cdot)$, *i.e.*,

$$D_c(\mathbf{v}_i, \mathbf{v}_j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{||\mathbf{v}_i|| ||\mathbf{v}_j||} = \frac{\sum_{k=1}^{n} v_i^k v_j^k}{\sqrt{\sum_{k=1}^{n} {v_i^k}^2} \cdot \sqrt{\sum_{k=1}^{n} {v_j^k}^2}}. \tag{5}$$

As Fig. 4 (left) shows, the sampled poses are more distinctive from each other when applying our FPS strategy.

Cross-distribution Sampling. Likewise, the synthetic grasping poses should also be novel and diverse. Simply blending them with real-world poses leads to uneven distribution, as shown in Fig. 4 (right), due to the presence of similar samples. While our FPS strategy can mitigate this issue, it can not ensure sufficient inclusion of real-world poses, since $M \ll N$, *e.g.*, $M = 10$ and $N = 500$ for HO3D. Additionally, the number of possible unique poses is unknown, making it challenging to define precise criteria for identifying different pose categories. Thus, we leverage the similarity relationship between the real-world and synthetic poses to adaptively control the sampling probability of the synthetic data. Specifically, for each object category, we calculate the similarity matrix and convert it into a discrete probability distribution P for the sampling process:

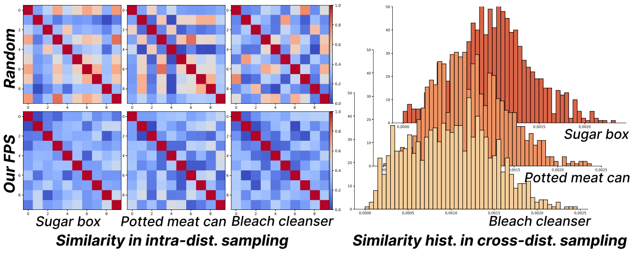

Figure 4. Visualizations in sampling. For three object categories in HO3D, we show the similarity of sampled real-world poses with our FPS or random sampling (left) and the histograms of similarity between sampled real-world and synthetic poses (right).

$$P(X = \mathbf{v}_i^s) = \text{Norm}(\sum_{j=1}^{M}(1 - D_c(\mathbf{v}_i^s, \mathbf{v}_j^r))), \tag{6}$$

where $\mathbf{v}_i^s \in \mathbf{Q}_s$ and $\mathbf{v}_j^r \in \mathbf{Q}_r$. $\text{Norm}(\cdot)$ denotes the min-max normalization. The sampling probability of a synthetic pose is inversely related to its similarity to all real-world poses, yielding a more balanced distribution.

3.4. Hand Mesh Reconstruction

In the task of 3D hand-mesh reconstruction from a single RGB image, the goal is to estimate the 3D hand pose and shape. To demonstrate the effectiveness of our approach, we adopt it to train two commonly-used pipelines, *i.e.*, MANO-based and vertex-based methods. The former represents the hand using a predefined template and utilizes MANO coefficients to control its pose and shape, which are regressed directly from the extracted features of the input image. The latter typically regresses the 3D coordinates of hand vertices in a coarse-to-fine manner. In this work, we select three recent methods as our baselines: HandOccNet (MANO-based), MobRecon (vertex-based), and H2ONet (vertex-based). They not only employ different hand-mesh representations but also utilize unique network architectures and scalable input resolutions. By applying our synthesized data to these baselines, we can effectively evaluate the generalizability and versatility of our approach.

	Methods	*Procrustes Alignment*						*Root-relative*					
		J-PE ↓	J-AUC ↑	V-PE ↓	V-AUC ↑	F@5 ↑	F@15 ↑	J-PE ↓	J-AUC ↑	V-PE ↓	V-AUC ↑	F@5 ↑	F@15 ↑
Multi-frame	S2Hand [13]	7.3	85.5	-	-	-	-	-	-	-	-	-	-
	VIBE [39]	6.4	87.1	-	-	-	-	-	-	-	-	-	-
	TCMR [15]	6.3	87.5	-	-	-	-	-	-	-	-	-	-
	H2ONet [66]	5.3	89.4	5.2	89.6	80.5	99.3	13.7	74.8	12.7	76.6	52.1	92.3
	Deformer [20]	5.2	89.6	-	-	-	-	-	-	-	-	-	-
Monocular	METRO [42]	7.0	-	-	-	-	-	15.2	-	-	-	-	-
	Spurr *et al.* [59]	6.8	86.4	-	-	-	-	17.3	69.8	-	-	-	-
	MeshGraphormer [43]	6.4	87.2	-	-	-	-	15.2	-	-	-	-	-
	Liu *et al.* [44]	6.6	-	-	-	-	-	15.3	-	-	-	-	-
	Tse *et al.* [60]	-	-	-	-	-	-	16.1	72.2	-	-	-	-
	HandOccNet [52]	5.8	88.4	5.5	89.0	78.0	99.0	14.0	74.8	13.1	76.6	51.5	92.4
	+ Our HandBooster	**5.2**	**89.6**	**5.0**	**89.9**	**81.3**	**99.2**	**11.9**	**77.8**	**11.5**	**78.5**	**55.6**	**93.6**
	MobRecon [12]	6.4	87.3	5.6	88.9	78.5	98.8	14.2	73.7	13.1	76.1	50.8	92.1
	+ Our HandBooster	**6.2**	**87.6**	**5.4**	**89.3**	**79.2**	**99.1**	**13.2**	**74.9**	**12.3**	**76.6**	**51.8**	**92.5**
	H2ONet [66]	5.7	88.9	5.5	89.1	80.1	99.0	14.0	74.6	13.0	76.2	51.3	92.1
	+ Our HandBooster	**5.1**	**89.8**	**5.1**	**89.8**	**81.3**	**99.2**	**12.9**	**76.0**	**12.5**	**76.5**	**52.1**	**92.2**

Table 1. Results on "S0" (default) data split of DexYCB. -: unavailable results. Our method ***consistently*** boosts all three baselines.

Methods	*Procrustes Alignment*						*Root-relative*					
	J-PE ↓	J-AUC ↑	V-PE ↓	V-AUC ↑	F@5 ↑	F@15 ↑	J-PE ↓	J-AUC ↑	V-PE ↓	V-AUC ↑	F@5 ↑	F@15 ↑
[52]	6.6	86.8	6.3	87.3	72.1	98.4	17.4	67.6	16.8	68.5	40.7	86.2
+ Ours	6.1 **+0.5**	87.8 **+1.0**	5.9 **+0.4**	88.3 **+1.0**	75.4 **+3.3**	98.7 **+0.3**	15.8 **+1.6**	70.4 **+2.8**	15.3 **+1.5**	71.3 **+2.8**	44.1 **+3.4**	88.5 **+2.3**
[12]	6.9	86.2	6.4	87.2	72.1	98.4	18.3	66.0	17.4	67.5	39.7	85.6
+ Ours	6.6 **+0.3**	86.7 **+0.5**	6.1 **+0.3**	87.9 **+0.7**	74.4 **+2.3**	98.6 **+0.2**	17.5 **+0.8**	67.7 **+1.7**	16.7 **+0.7**	69.1 **+1.6**	41.1 **+1.4**	86.8 **+1.2**
[66]	6.2	87.7	6.4	87.2	72.4	98.6	17.6	67.9	17.2	68.4	39.4	86.2
+ Ours	6.0 **+0.2**	88.1 **+0.4**	6.0 **+0.4**	88.1 **+0.9**	75.4 **+3.0**	98.7 **+0.1**	17.0 **+0.6**	68.9 **+1.0**	16.4 **+0.8**	69.7 **+1.3**	41.5 **+2.1**	87.2 **+1.0**

Table 2. Results on "S1" (unseen subjects) data split of DexYCB. Our method ***consistently*** improves all baselines for all metrics.

4. Experiments

4.1. Experimental Settings

Datasets. We employ the following commonly-used hand-object benchmark datasets in our experiments. (i) **DexYCB** [8] (a real-world dataset): we use the default "S0" split (train/test: 406,888/78,768 samples) and the more challenging "S1" split with unseen subjects (train/test: 7/2 subjects). (ii) **HO3D** (version 2) [22] (a real-world dataset): note that evaluation can only be done by submitting results to the official server. (iii) **ObMan** [23] (a rendering-based synthetic dataset): we select it as the competitor of our generated data. Note that other rendering/generative-based candidates are not available, including YCB-Afford [17] (download link not accessible), ArtiBoost [71] (data not open-sourced), and HOGAN [31] (missing essential files for training). To show the improvement in generalizability, we additionally use **MOW** [7] as the in-the-wild test set, which provides 512 annotated samples.

Evaluation Metrics. We adopt common metrics following [8, 12, 22, 52, 66]. J-PE/V-PE denotes the root-relative joint/vertex position error, measuring the average Euclidean distance in mm between the predicted and ground-truth 3D hand joint/vertex coordinates. J-AUC/V-AUC computes the area under the percentage of correct keypoints (PCK) curve in several error thresholds for joint/vertex. F@5 and F@15 measure the harmonic mean of the recall and precision for vertex with $5mm$ and $15mm$ thresholds. Procrustes Alignment (PA) aligns the orientation, translation, and scale of the estimated results to match the ground truths. Fréchet Inception Distance [26] (FID) measures the image fidelity.

Implementation Details. Adam optimizer [38] is applied to train the diffusion model and the reconstruction methods. 2D conditions are 256×256 and rendered using PyRender. The diffusion model is trained on 16 NVIDIA 2080Ti GPUs with a batch size of 256 and a learning rate of 0.00008 for 700,000/200,000 iterations on DexYCB/HO3D. Please refer to our supp. material for other details.

4.2. Comparison with State-of-the-art Methods

Evaluation on DexYCB. We first conduct quantitative comparisons on DexYCB. To demonstrate the effectiveness of our method, we evaluate metrics before and after performing PA, as presented in Tab. 1 and Tab. 2 for "S0" and "S1" data splits, respectively. Additionally, Fig. 5 (left) visualizes the root-relative mesh PCK/AUC comparison for the more challenging "S1" split. Our synthetic data significantly improves the performance of HandOccNet [52] and the monocular-based H2ONet [66], enabling them even surpass some multi-frame methods, such as Deformer [20]. MobRecon [12] exhibits relatively small performance gains due to its mobile-friendly designs. Its capability is limited

	Methods	J-PE ↓	J-AUC ↑	V-PE ↓	V-AUC ↑	F@5 ↑	F@15 ↑
Multi-frame	Hasson *et al.* [24]	11.4	77.3	11.4	77.3	42.8	93.2
	Hasson *et al.* [25]	-	-	14.7	-	39.0	88.0
	S2Hand [13]	11.4	77.3	11.2	77.7	45.0	93.0
	Liu *et al.* [44]	9.8	-	9.4	81.2	53.0	95.7
	VIBE [39]	9.9	-	9.5	-	52.6	95.5
	TCMR [15]	11.4	-	10.9	-	46.3	93.3
	TempCLR [80]	10.6	-	10.6	-	48.1	93.7
	Deformer [20]	9.4	-	9.1	-	54.6	96.3
	H2ONet [66]	8.5	82.9	8.6	82.8	57.0	96.6
Monocular	Pose2Mesh [14]	12.5	-	12.7	-	44.1	90.9
	I2L-MeshNet [50]	11.2	-	13.9	-	40.9	93.2
	ObMan [23]	11.1	-	11.0	77.8	46.0	93.0
	HO3D [22]	10.7	78.8	10.6	79.0	50.6	94.2
	METRO [42]	10.4	-	11.1	-	48.4	94.6
	Liu *et al.* [44]	10.2	79.7	9.8	80.4	52.9	95.0
	I2UV-HandNet [10]	9.9	80.4	10.1	79.9	50.0	94.3
	Tse *et al.* [60]	-	-	10.9	-	48.5	94.3
	AMVUR [36]	8.3	83.5	8.2	83.6	60.8	96.5
	HandOccNet [52]	9.1	81.9	9.0	81.9	56.1	96.2
	HandOccNet* [52]	9.6	80.8	9.6	80.7	52.4	95.4
	+ Our HandBooster	**8.5**	**82.9**	**8.6**	**82.9**	**57.7**	**97.2**
	MobRecon [12]	9.4	81.3	9.5	81.0	53.3	95.5
	MobRecon† [12]	9.2	81.6	9.4	81.2	53.8	95.7
	+ Our HandBooster	**8.7**	**82.6**	**8.8**	**82.5**	**56.1**	**97.0**
	H2ONet [66]	9.0	82.0	9.0	81.9	55.4	96.0
	+ Our HandBooster	**8.2**	**83.6**	**8.4**	**83.2**	**58.5**	**97.2**

Table 3. Results on HO3D (*Procrustes Alignment*). *: reproduced results using its official code. †: complement data is used. -: unavailable results. Our method brings improvement consistently.

by its lower FLOPs (0.46G) and the number of parameters (8.23M) compared to H2ONet (0.74G/25.88M) and HandOccNet (15.47G/37.22M). Our HandBooster consistently boosts the performance of all three baselines across all metrics and two data splits, giving strong evidence of its effectiveness. It even outperforms multi-frame methods (see Tab. 1), by taking only a single view as its input.

Fig. 6 (a-b) and (d) show qualitative results on DexYCB. Note that we directly use models pre-trained on DexYCB's "S0" split to test on MOW. Comparing the results w/ and w/o using our data, the performance gains clearly show HandBooster's efficacy and generalizability. Fig. 7 shows some of the generated samples with novel views/poses, where the fine-grained alignment with conditions and the realistic appearance show the controllability of HandBooster. More results are shown in the supp. material.

Evaluation on HO3D. We perform the same experiments on HO3D. Since the ground truths of the test set are not publicly available, some results are from previous papers and the official evaluation server. Tabs. 3 and 4 show results before and after PA, respectively. To facilitate comparison, we also provide the mesh PCK/AUC after PA, visualized in Fig. 5 (right). It is clear that all our baselines consistently achieve improved precision on all the metrics when boosted by our generated data, showing the great effectiveness and robustness of our method. For an intuitive comparison, we present visual comparisons before and after applying our method in Fig. 6 (c). Benefiting from the diverse and novel samples in our generated data, all these baselines become able to produce plausible shapes and estimate accurate hand orientations, even under severe occlusions in the inputs.

Methods	J-PE ↓	J-AUC ↑	V-PE ↓	V-AUC ↑	F@5 ↑	F@15 ↑
[52]	24.9	53.9	24.2	55.1	26.0	72.9
[52]*	25.1	53.3	24.5	54.4	25.6	72.8
+ Ours	21.1 **+4.0**	59.4 **+6.1**	20.5 **+4.0**	60.4 **+6.0**	28.7 **+3.1**	77.9 **+5.1**
[12]	25.2	53.7	24.4	55.0	26.4	72.0
+ Ours	23.4 **+1.8**	56.5 **+2.8**	22.6 **+1.8**	57.9 **+2.9**	27.7 **+1.3**	75.3 **+3.3**
[66]	26.3	52.3	25.5	53.5	24.9	71.5
+ Ours	24.0 **+2.3**	56.7 **+4.4**	23.3 **+2.2**	57.7 **+4.2**	26.6 **+1.7**	74.4 **+2.9**

Table 4. Results on HO3D (*Root-relative*). *: reproduced results using its official code. Our method boosts baselines significantly.

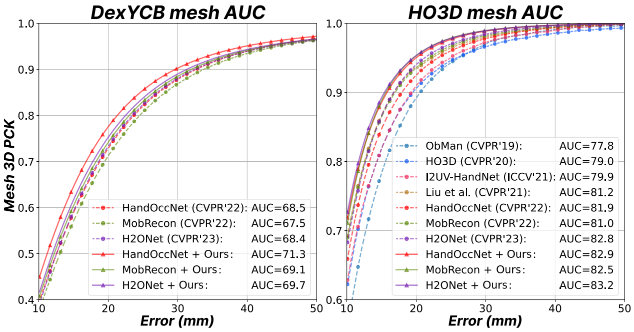

Figure 5. The mesh AUC comparison under different thresholds. All baselines are improved consistently after applying our data.

4.3. Ablation Studies

We conduct ablation studies on DexYCB to evaluate the effectiveness of HandBooster, as shown in Tab. 5. MobRecon is selected as the baseline due to its fast training speed.

Rendering-based Synthesis. We show the importance of realism in data generation by comparing our data with ObMan, where hand appearance, grasping pose, and background are randomly selected for rendering. Comparing Rows (a) and (b), introducing ObMan only slightly improves the root-relative metrics after PA. Correspondingly, comparing Rows (a) and (f), our data enhances performance on all metrics even without the adoption of our other techniques, revealing the importance of realism in the data generation. Further, we compute the FID score between our data/ObMan and the training set of DexYCB. The scores of 2.91/8.52 reflect the realism of our generated data.

Content-aware Conditions. To better understand the influence of different conditions for 3D hand-mesh reconstruction, we evaluate various candidate combinations. Due to the lack of shape information, the skeleton and foreground segmentation are not used in this analysis. The results, shown in Rows (d-f) of Tab. 5, reveal that incorporating both normal and texture yields the highest performance. Substituting either component with alternative candidates leads to a significant decrease in performance, thus supporting our idea of constructing conditions to be both informative and easy to understand. Further, by introducing embedded hand orientation as an additional control, our model shows noticeable improvements in root-relative metrics, further validating the effectiveness of our design.

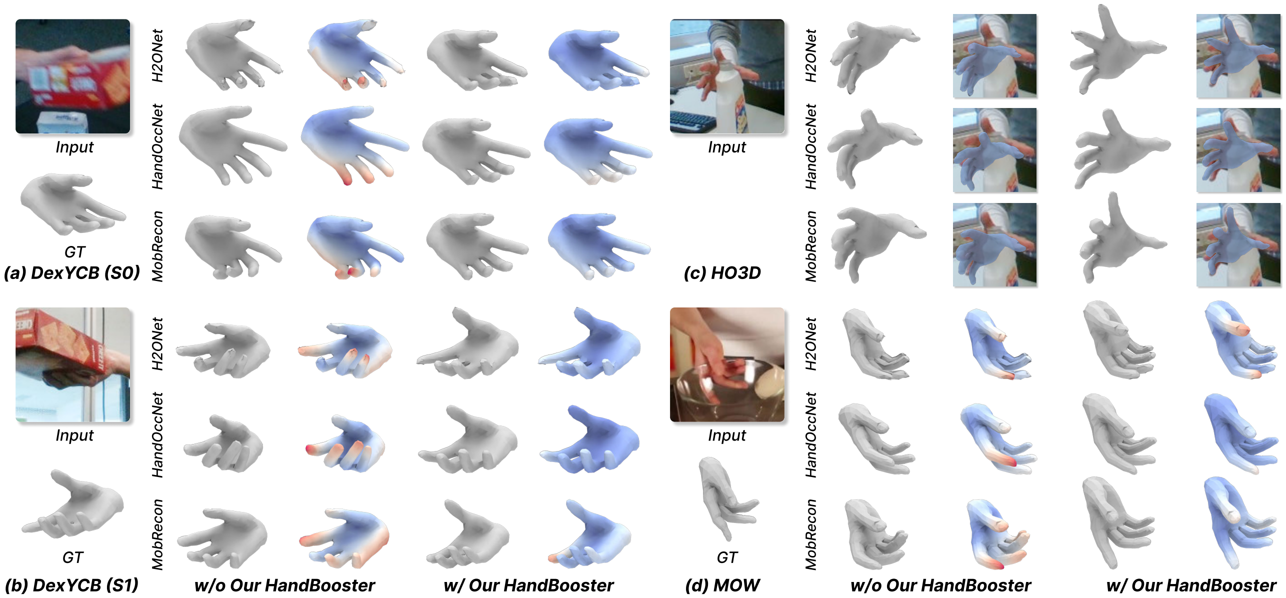

Figure 6. Qualitative comparison of [12, 52, 66] w/ and w/o our HandBooster. (For each case, left: predicted results, right: error map.)

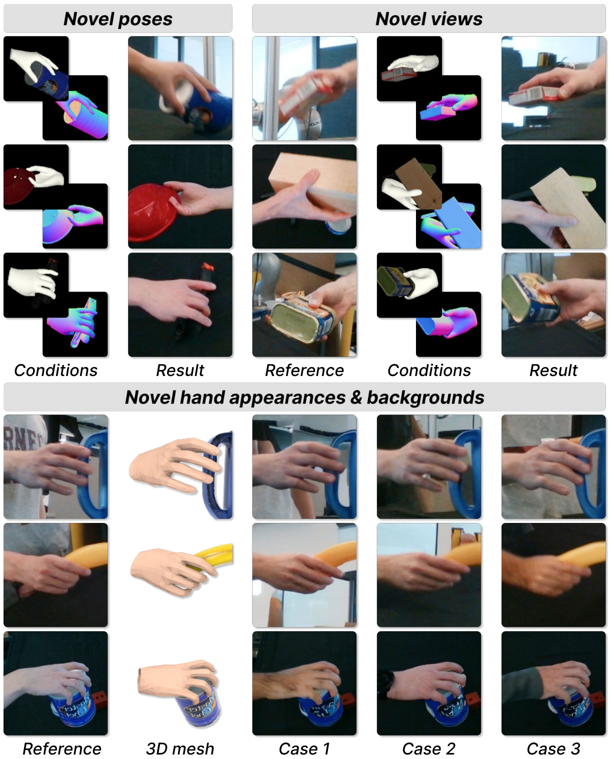

Figure 7. Generated examples for novel views/poses with realistic hand appearance and backgrounds.

Novel Condition Creator. We also investigate the impact of components in our novel condition creator. Comparing Rows (f) and (g), though only utilizing novel grasping poses boosts root-relative performance, certain poses may dominate the entire distribution and limit the performance. Comparing Rows (h-i) with (g), performing the intra- and cross-distribution sampling sequentially brings boosts across all metrics clearly, demonstrating their necessities. Further experimental results are presented in the supp. material.

	Models	*Root-relative*		*Procrustes Align.*	
		J-PE ↓	V-PE ↓	J-PE ↓	V-PE ↓
(a)	Baseline	14.20	13.05	6.36	5.59
(b)	w/ ObMan	14.00	13.00	6.40	5.60
(c)	w/ Depth & Segment.	14.91	13.94	6.64	5.85
(d)	w/ Normal & Segment.	14.82	13.85	6.60	5.81
(e)	w/ Normal & Texture	13.93	12.92	6.37	5.53
(f)	+ Embedded Orientation	13.79	12.78	6.33	5.49
(g)	+ Novel Conditions	13.51	12.57	6.34	5.55
(h)	+ Intra-dist. Sampling	13.42	12.42	6.28	5.46
(i)	+ Cross-dist. Sampling	**13.25**	**12.34**	**6.20**	**5.36**

Table 5. Ablation study on major components.

5. Conclusion

We presented HandBooster, a new generative method to boost 3D hand-mesh reconstruction by enhancing the data diversity. First, we create a conditional generative space, from which we can controllably produce realistic and diverse hand-object images with reliable 3D annotations. Then, we explore this space to produce novel and diverse training samples by formulating a novel condition creator and two similarity-aware sampling strategies. Extensive experiments on three baselines and two common benchmarks demonstrate our effectiveness and SOTA performance.

Acknowledgments This work was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. T45-401/22-N and No. CUHK 14201921) and the National Natural Science Foundation of China under grant No. 62372091. Hao Xu thanks for the care & support from Yutong Zhang & his family.

References

[1] Autodesk, INC. Maya. https:/autodesk.com/maya, 2018. 1, 2

[2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for RGB-based dense 3D hand pose estimation via neural rendering. In *CVPR*, pages 1067–1076, 2019. 2

[3] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Weakly-supervised domain adaptation via GAN and mesh model for estimating 3D hand poses interacting objects. In *CVPR*, pages 6121–6131, 2020. 2

[4] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-DPM: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. *arXiv preprint arXiv:2201.06503*, 2022. 2

[5] Adnane Boukhayma, Rodrigo de Bem, and Philip H.S. Torr. 3D hand shape and pose from images in the wild. In *CVPR*, pages 10843–10852, 2019. 2

[6] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. The YCB object and model set: Towards common benchmarks for manipulation research. In *ICRA*, pages 510–517. IEEE, 2015. 4

[7] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *ICCV*, pages 12417–12426, 2021. 2, 6

[8] 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 *CVPR*, pages 9044–9053, 2021. 1, 4, 6

[9] Liangjian Chen, Shih-Yao Lin, Yusheng Xie, Yen-Yu Lin, and Xiaohui Xie. Temporal-aware self-supervised learning for 3D hand pose and mesh estimation in videos. In *WACV*, pages 1050–1059, 2021. 2

[10] 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 *ICCV*, pages 12929–12938, 2021. 7

[11] 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*, pages 13274–13283, 2021. 2

[12] 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 *CVPR*, pages 20544–20554, 2022. 2, 6, 7, 8

[13] 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 *CVPR*, pages 10451–10460, 2021. 2, 6, 7

[14] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2Mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. In *ECCV*, pages 769–787, 2020. 7

[15] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3D human pose and shape from a video. In *CVPR*, pages 1964–1973, 2021. 6, 7

[16] Blender Online Community. Blender. http://www.blender.org, 2019. 1, 2

[17] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. GANHand: Predicting human grasp affordances in multi-object scenes. In *CVPR*, pages 5031–5041, 2020. 1, 2, 6

[18] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2

[19] Yuval Eldar, Michael Lindenbaum, Moshe Porat, and Yehoshua Y Zeevi. The farthest point strategy for progressive image sampling. *IEEE TIP*, 6(9):1305–1315, 1997. 5

[20] Qichen Fu, Xingyu Liu, Ran Xu, Juan Carlos Niebles, and Kris M Kitani. Deformer: Dynamic fusion transformer for robust hand pose estimation. *arXiv preprint arXiv:2303.04991*, 2023. 2, 6, 7

[21] 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*, pages 10833–10842, 2019. 2

[22] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A method for 3D annotation of hand and object poses. In *CVPR*, pages 3196–3206, 2020. 1, 4, 6, 7

[23] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, pages 11807–11816, 2019. 1, 2, 6, 7

[24] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *CVPR*, pages 571–580, 2020. 7

[25] Yana Hasson, Gül Varol, Cordelia Schmid, and Ivan Laptev. Towards unconstrained joint hand-object reconstruction from RGB videos. In *3DV*, pages 659–668, 2021. 7

[26] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 6

[27] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 3

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 2, 3

[29] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. Simple diffusion: End-to-end diffusion for high resolution images. *arXiv preprint arXiv:2301.11093*, 2023. 2

[30] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 2

[31] Hezhen Hu, Weilun Wang, Wengang Zhou, and Houqiang Li. Hand-object interaction image generation. In *NeurIPS*, pages 23805–23817, 2022. 1, 2, 6

[32] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5D heatmap regression. In *ECCV*, pages 118–134, 2018. 2

[33] Juntao Jian, Xiuping Liu, Manyi Li, Ruizhen Hu, and Jian Liu. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In *ICCV*, pages 14713–14724, 2023. 2

[34] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *ICCV*, pages 11107–11116, 2021. 2

[35] Hai Jiang, Ao Luo, Haoqiang Fan, Songchen Han, and Shuaicheng Liu. Low-light image enhancement with wavelet-based diffusion models. *ACM TOG*, 42(6):1–14, 2023. 2

[36] Zheheng Jiang, Hossein Rahmani, Sue Black, and Bryan M Williams. A probabilistic attention model with occlusion-aware texture regression for 3d hand reconstruction from a single rgb image. In *CVPR*, pages 758–767, 2023. 7

[37] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022. 2, 4

[38] P. Diederik Kingma and Lei Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 3, 6

[39] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, pages 5253–5263, 2020. 6, 7

[40] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M. Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *CVPR*, pages 4990–5000, 2020. 1, 2

[41] Haipeng Li, Hai Jiang, Ao Luo, Ping Tan, Haoqiang Fan, Bing Zeng, and Shuaicheng Liu. DMHomo: Learning homography with diffusion models. *ACM TOG*, 2024. 2

[42] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, pages 1954–1963, 2021. 2, 6, 7

[43] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, pages 12939–12948, 2021. 2, 6

[44] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3D hand-object poses estimation with interactions in time. In *CVPR*, pages 14687–14697, 2021. 2, 6, 7

[45] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 289–299, 2023. 2

[46] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022. 2

[47] Matthew Matl. Pyrender. `https://github.com/mmatl/pyrender`, 2019. 1

[48] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *CVPR*, pages 4460–4470, 2019. 2

[49] Andrew T Miller and Peter K Allen. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 11(4):110–122, 2004. 2

[50] 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 *ECCV*, pages 752–768, 2020. 2, 7

[51] 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*, pages 548–564, 2020. 2

[52] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. HandOccNet: Occlusion-robust 3D hand mesh estimation network. In *CVPR*, pages 1496–1505, 2022. 2, 6, 7, 8

[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 3

[54] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 2

[55] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. pages 2256–2265, 2015. 2

[56] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[57] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *NeurIPS*, 32, 2019. 2

[58] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[59] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3D hand pose estimation via biomechanical constraints. In *ECCV*, pages 211–228, 2020. 6

[60] 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 *CVPR*, pages 1664–1674, 2022. 2, 6, 7

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4

[62] Weikang Wan, Haoran Geng, Yun Liu, Zikang Shan, Yaodong Yang, Li Yi, and He Wang. Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. *arXiv preprint arXiv:2304.00464*, 2023. 2

[63] Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. In *ICRA*, pages 11359–11366. IEEE, 2023. 2, 4

[64] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *ICLR*, 2023. 2

[65] Yilin Wen, Hao Pan, Lei Yang, Jia Pan, Taku Komura, and Wenping Wang. Hierarchical temporal transformer for 3D hand pose estimation and action recognition from egocentric RGB videos. In *CVPR*, pages 21243–21253, 2023. 2

[66] Hao Xu, Tianyu Wang, Xiao Tang, and Chi-Wing Fu. H2onet: Hand-occlusion-and-orientation-aware network for real-time 3d hand mesh reconstruction. In *CVPR*, pages 17048–17058, 2023. 2, 6, 7, 8

[67] Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, et al. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *CVPR*, pages 4737–4746, 2023. 2, 4

[68] John Yang, Hyung Jin Chang, Seungeui Lee, and Nojun Kwak. Seqhand: RGB-sequence-based 3D hand pose and shape estimation. In *ECCV*, pages 122–139, 2020. 2

[69] Lixin Yang, Jiasen Li, Wenqiang Xu, Yiqun Diao, and Cewu Lu. BiHand: Recovering hand mesh with multi-stage bisected hourglass networks. In *BMVC*, 2020. 2

[70] Linlin Yang, Shicheng Chen, and Angela Yao. SemiHand: Semi-supervised hand pose estimation with consistency. In *ICCV*, pages 11364–11373, 2021. 2

[71] Lixin Yang, Kailin Li, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In *CVPR*, pages 2750–2760, 2022. 1, 2, 6

[72] Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *ICCV*, pages 19717–19728, 2023. 2

[73] Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *CVPR*, pages 22479–22489, 2023. 2

[74] Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang. Interacting two-hand 3D pose and shape reconstruction from single color image. In *ICCV*, pages 11354–11363, 2021. 2

[75] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 2

[76] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular RGB image. In *ICCV*, pages 2354–2364, 2019. 2

[77] 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*, pages 11281–11292, 2021.

[78] Zimeng Zhao, Xi Zhao, and Yangang Wang. TravelNet: Self-supervised physically plausible hand motion learning from monocular color images. In *ICCV*, pages 11666–11676, 2021.

[79] 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 *CVPR*, pages 5346–5355, 2020. 2

[80] Andrea Ziani, Zicong Fan, Muhammed Kocabas, Sammy Christen, and Otmar Hilliges. TempCLR: Reconstructing hands via time-coherent contrastive learning. In *3DV*, pages 627–636. IEEE, 2022. 7

[81] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In *ICCV*, pages 813–822, 2019. 2