arXiv:2503.13303v1 [cs.CV] 17 Mar 2025

UniHOPE: A Unified Approach for Hand-Only and Hand-Object Pose Estimation

Yinqiao Wang* Hao Xu* Pheng-Ann Heng Chi-Wing Fu
Department of Computer Science and Engineering
Institute of Medical Intelligence and XR
The Chinese University of Hong Kong
{yqwang,xuhao,pheng,cwfu}@cse.cuhk.edu.hk

Abstract

Estimating the 3D pose of hand and potential hand-held object from monocular images is a longstanding challenge. Yet, existing methods are specialized, focusing on either bare-hand or hand interacting with object. No method can flexibly handle both scenarios and their performance degrades when applied to the other scenario. In this paper, we propose UniHOPE, a unified approach for general 3D hand-object pose estimation, flexibly adapting both scenarios. Technically, we design a grasp-aware feature fusion module to integrate hand-object features with an object switcher to dynamically control the hand-object pose estimation according to grasping status. Further, to uplift the robustness of hand pose estimation regardless of object presence, we generate realistic de-occluded image pairs to train the model to learn object-induced hand occlusions, and formulate multi-level feature enhancement techniques for learning occlusion-invariant features. Extensive experiments on three commonly-used benchmarks demonstrate UniHOPE's SOTA performance in addressing hand-only and hand-object scenarios. Code will be released on https://github.com/JoyboyWang/UniHOPE_Pytorch.

1. Introduction

Estimating the 3D pose of hand and potential hand-held objects from monocular images is a long-standing task with applications in VR/AR, human-computer interactions, *etc*.

However, existing methods are divided. As Fig. 1 illustrates, hand pose estimation (HPE) methods [5, 9, 25, 36, 39, 49, 64] predict the 3D hand pose without considering the hand-held object. Conversely, hand-object pose estimation (HOPE) methods [18, 19, 31, 32, 47] assume the presence of a hand-held object and perform object pose estimation with an extra object branch. Yet, they always make predictions even there is no object. Neither approach offers the flexibility to consider both hand-only and hand-object scenarios.

Tab. 1 provides a detailed analysis of the performance of state-of-the-art (SOTA) HPE methods [39, 49, 64] and HOPE methods [18, 31]. We observe an obvious performance degradation when these methods are applied across different scenes (see "*Hand-Only↔Hand-Object Scene*") due to their task-specific designs. Though training on all scenes helps, it negatively impacts their original task performance (see "*All→Hand-Only/Hand-Object Scene*"), revealing their limited generalization capabilities. This observation motivates the need for a unified approach that can adapt effectively to both hand-only and hand-object scenes.

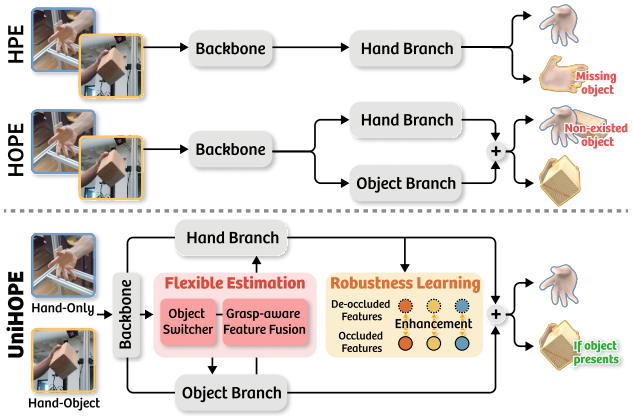


Figure 1. Existing approaches (top) for 3D hand pose estimation are either Hand Pose Estimation (HPE), which predicts hand pose only, or Hand-Object Pose Estimation (HOPE), which assumes hand-held object. Our novel UniHOPE approach (bottom) offers flexibility and robustness to handle both scenes in a unified manner.

In this work, we present UniHOPE, the first method to unify HPE and HOPE by addressing (i) basic criteria: *adaptively switch between two scenes*; and (ii) advanced criteria: *robustly estimate hand pose regardless of object presence*.

First, to meet the basic criteria, we propose that the hand pose must always be predicted regardless of whether the hand is grasping an object or not, while the object pose should be estimated only if the object is present. Though a straightforward solution is to manually select existing SOTA HPE and HOPE methods according to the input scene, this approach is suboptimal, as switching between models leads to incoherent results and prevents joint optimization for one

* Equal contribution.

HPE	*Hand-Only Scene*		*Hand-Only → Hand-Object Scene*		*All → Hand-Only Scene*		*All → Hand-Object Scene*	
	J-PE ↓	V-PE ↓	J-PE ↓	V-PE ↓	J-PE ↓	V-PE ↓	J-PE ↓	V-PE ↓
HandOccNet [39]	12.98	12.52	19.60 (**-6.62**)	18.95 (**-6.43**)	13.16 (**-0.18**)	12.70 (**-0.18**)	14.58	14.10
H2ONet [49]	13.34	13.13	21.98 (**-8.64**)	21.42 (**-8.30**)	14.14 (**-0.80**)	14.00 (**-0.87**)	15.20	15.03
SimpleHand [64]	14.05	13.51	18.37 (**-4.32**)	17.54 (**-4.03**)	14.63 (**-0.58**)	13.96 (**-0.45**)	14.88	14.21
HOPE	*Hand-Object Scene*		*Hand-Object → Hand-Only Scene*		*All → Hand-Object Scene*		*All → Hand-Only Scene*	
	J-PE ↓	V-PE ↓	J-PE ↓	V-PE ↓	J-PE ↓	V-PE ↓	J-PE ↓	V-PE ↓
Keypoint Trans. [18]	17.99	17.57	25.10 (**-7.11**)	24.40 (**-6.83**)	18.79 (**-1.00**)	18.35 (**-0.78**)	19.75	19.26
HFL-Net [31]	14.61	14.13	19.39 (**-4.78**)	18.61 (**-4.48**)	14.77 (**-0.16**)	14.29 (**-0.16**)	13.61	13.10

Table 1. Existing HPE methods trained on hand-only scene exhibit obvious performance degradation when testing on hand-object scene (1st *vs*. 2nd columns). Though training on all scenes (3rd & 4th columns) helps to improve metrics on the hand-object scene (2nd *vs*. 4th columns), their original performance is adversely affected (1st *vs*. 3rd columns). The HOPE methods also exhibit a similar pattern (see bottom part). These results demonstrate the inabilities of the existing methods to flexibly handle hand-only and hand-object scenes altogether.

model to work on both scenes. Importantly, we design an end-to-end method that dynamically controls object pose estimation through an internal object switcher by estimating the confidence of the grasping status, thereby promoting compatibility with various scenarios. However, solely adopting existing network architectures along with our object switcher encounters another issue caused by the commonly-used hand-object information interaction structure [31, 32, 45]. When no object is present, extracting object features is unnecessary, so irrelevant object-to-hand feature transitions compromise hand pose estimation accuracy. To overcome this issue, we formulate a grasp-aware feature fusion module to utilize grasping confidence to select effective object features in hand-object feature fusion.

Second, the advanced criteria emphasize coherent and robust hand pose estimation, regardless of whether the hand is grasping an object or not. As hand-held objects frequently cause severe occlusions, it is essential to learn occlusion-invariant features to accurately recover hand poses. Ideally, the feature of a non-occluded hand serves as the optimal representation for an occluded hand with the same pose, as it simplifies the prediction difficulty. To facilitate this, we propose training the model by transferring knowledge from the corresponding non-occluded hands to the occluded ones. Due to the scarcity of such paired data, we innovatively leverage diffusion-based generative models to create realistic de-occluded hand images from the originally-occluded ones. Subsequently, we adopt multi-level feature enhancement techniques to help the network simulate occlusion-invariant features by utilizing information from the de-occluded hand images in a self-distillation framework.

Our main contributions are summarized as follows:

- We demonstrate the necessity for a unified solution to hand-object pose estimation and propose UniHOPE, a novel approach to handle general hand-object scenarios.
- We design an internal object switcher that provides flexibility across different scenes and formulate a grasp-aware feature fusion module to adaptively utilize the effective object information based on grasping status.
- We propose an occlusion-invariant feature learning strategy for robustness, first using a generative de-occluder to prepare paired de-occluded hand images and then applying feature enhancement at multiple levels.
- Extensive experiments on three widely-used datasets in our unified setting show the SOTA performance of UniHOPE.

2. Related Work

Monocular 3D Hand Pose Estimation (HPE). Most methods formulate HPE by regressing MANO coefficients [2–4, 10, 51, 57, 60–63, 65]. Other common representations include voxels [26, 36, 37, 52], implicit functions [25], and meshes [8, 9, 29, 49, 64]. While they achieve superior performance in predicting hand poses, they do not account for object pose estimation during hand-object interactions, which is critical for practical applications.

Monocular 3D Hand-Object Pose Estimation (HOPE). To jointly estimate the hand and object poses simultaneously, recent works can be categorized into two main streams: (i) Template-free methods reconstruct objects without knowing their 3D models. Hasson *et al*. [19] recover the object mesh from a deformed icosphere. Tse *et al*. [45] propose to optimize hand and object meshes iteratively. More recent works leverage implicit [11, 12, 24, 54, 56] or neural fields [13] to represent the hand and object; yet, these methods often struggle to accurately model unseen objects due to limited data prior. (ii) Template-based methods assume the 3D object model is known, focusing on regressing its pose. Liu *et al*. [32] proposes a semi-supervised learning approach. Lin *et al*. [31] design a dual-branch backbone to leverage mutual hand-object information. Though these methods produce promising results, they focus on the hand-object scene, lacking the flexibility to handle the hand-only scenario.

Hand-Object Image Synthesis. Since we generate paired de-occluded hand images for unified scenarios, we also review methods for hand-object image synthesis. Rendering-based methods [15, 19, 53] use common tools [1, 14, 35] to render images as augmented data for HPE and HOPE tasks. With the advance of generative models [16, 21, 42–44, 48], recent methods generate more realistic data for various pur-

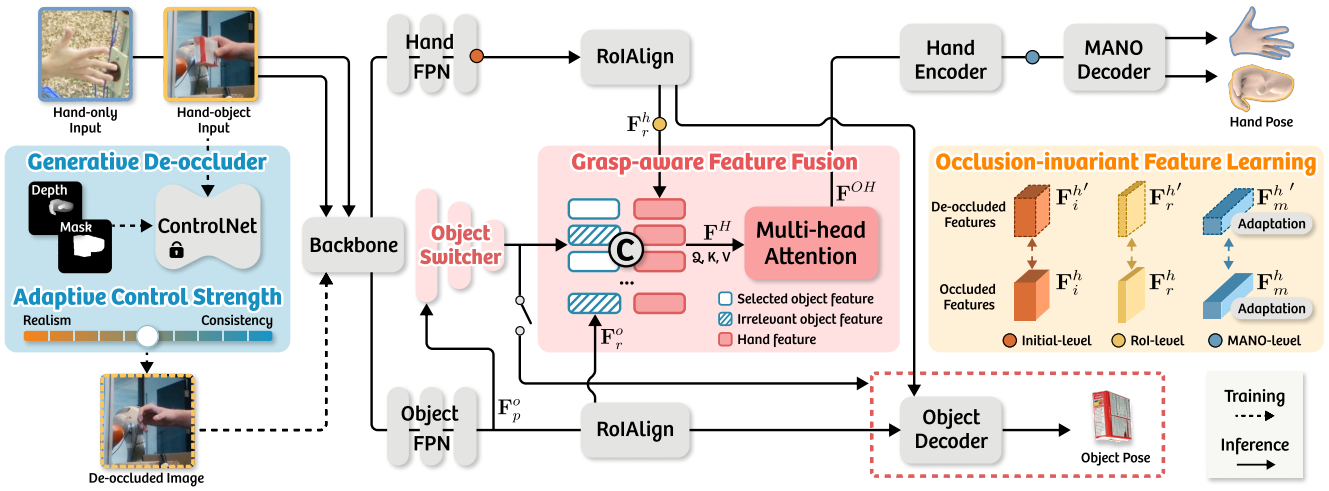


Figure 2. Our UniHOPE framework. (i) We first de-occlude hand images occluded by objects to form pairs, conditioned on the depth map and hand-object mask, with adaptive selection of control strength to produce high-quality samples; (ii) to accommodate both hand-only and hand-object scenes, our object switcher dynamically controls the object output by predicting grasping status, which guides the feature fusion module to eliminate irrelevant object features; and (iii) to robustly estimate hand pose, our multi-level feature enhancement techniques utilize paired data to learn occlusion-invariant hand features.

poses, including image generation [22, 38, 59], object-prior guidance [55], hand re-malformation [33], and training data augmentation [40, 50]. In this work, we propose to generate paired de-occluded hand images to facilitate learning occlusion-invariant hand features, thereby improving robustness to handle both hand-only and hand-object scenes.

3. Method

3.1. Overview

Distinct from previous studies, we address a more general scenario, *i.e.*, the model always predicts hand pose, regardless of whether the hand is grabbing an object or not. If an object is present, our model also estimates its pose.

Fig. 2 shows our UniHOPE pipeline. First, we propose to dynamically estimate hand-object pose in an end-to-end manner (see the red section in Fig. 2), where our object switcher flexibly controls the output and our grasp-aware feature fusion module integrates grasp-relevant object information (Sec. 3.2). Next, to improve robustness against occlusions, we first design our generative de-occluder to prepare high-quality paired data by adaptively adjusting control strengths (see blue section), which is then used to learn occlusion-invariant features through our multi-level feature enhancement techniques (see yellow section, Sec. 3.3). Finally, we detail the loss functions used in our approach in Sec. 3.4.

3.2. Dynamic Hand-Object Pose Estimation

To accommodate both HPE and HOPE, it is essential to consistently predict hand poses while estimating object poses only when an object is present. Directly combining existing HPE and HOPE methods is insufficient due to the incoherence introduced by model switching and the lack of joint optimization. In this section, we present our end-to-end dynamic hand-object pose estimation approach. To flexibly control the object output, we introduce an object switcher that predicts grasping status, trained with automatically-generated labels. Further, we propose a grasp-aware feature fusion module guided by the grasping status to prevent object-to-hand irrelevant feature transitions when no object is present.

Grasping Label Preparation. To support model training, we automatically prepare grasping status labels, as existing datasets [6, 17] do not provide this information. Following [50], we compute the isotropic Relative Rotation Error (RRE) and Relative Translation Error (RTE) between the object poses in the initial and current frames:

$$\text{RRE} = \arccos\left(\frac{\text{trace}({\xi_R^t}^\top \xi_R^0 - 1)}{2}\right), \ \text{RTE} = ||\xi_T^t - \xi_T^0||_2, \quad (1)$$

where $\xi_R \in \mathbb{R}^{3\times3}$ and $\xi_T \in \mathbb{R}^3$ denote the object rotation matrix and translation vector, respectively. The superscripts 0 and t indicate the frame index. The object is labeled as grasped if the computed errors exceed a defined threshold.

Object Switcher. With the prepared labels, we employ a multi-layer perceptron (MLP) $g(\cdot)$ to predict the grasping status from the object feature $\mathbf{F}_p^o$, which is extracted by the Feature Pyramid Network (FPN) [30] from the input image. This process is supervised by the binary cross-entropy loss:

$$\mathcal{L}^s = -\sum_{j=0}^{1} \mathbb{1}(\hat{G} = j) \cdot \log \frac{\exp(g(\mathbf{F}_p^o)_j)}{\sum_{k=0}^{1} \exp(g(\mathbf{F}_p^o)_k)}, \quad (2)$$

where $\hat{G}$ is the ground-truth grasping label and $\mathbb{1}(\cdot)$ is the indicator function. During testing, the object pose estimation branch is deactivated if predicted as non-grasping, providing more accurate responses for hand-object interactions.

Grasp-aware Feature Fusion. Previous studies have shown that feature interaction between hand and object can effectively enhance performance in hand-object scenes [31, 32, 45]; however, such interaction can disrupt hand feature learning in the hand-only scene due to the absence of objects (see Tab. 1). To mitigate interference from irrelevant object features, we design the grasp-aware feature fusion. During training, the object feature $\mathbf{F}_r^o$ and the hand feature $\mathbf{F}_r^h$ produced by RoIAlign [20] are concatenated to form the feature $\mathbf{F}^H$ only when the object is predicted as grasped. Next, $\mathbf{F}^H$ is processed through a multi-head attention block [46], resulting in the fused hand-object feature $\mathbf{F}^{OH}$:

$$\begin{aligned} \mathbf{F}^H &= \text{Concat}(\mathbf{F}_r^h, s\mathbf{F}_r^o + (1-s)\mathbf{F}_r^h) \quad \text{and} \\ \mathbf{F}^{OH} &= \text{Softmax}(\frac{\mathbf{F}^H \mathbf{F}^{H^T}}{\sqrt{d^H}})\mathbf{F}^H, \end{aligned} \tag{3}$$

where s=Argmax$(g(\mathbf{F}_p^o))$ indicates the predicted grasping status. Concat$(\cdot)$ and Softmax$(\cdot)$ represent the concatenation and soft-max operations along the channel dimension, respectively. d^H is the channel dimension of $\mathbf{F}^H$. This approach enables the network to flexibly toggle object outputs while maintaining robust feature representations for the hand across various scenes. Then, similar to [31], $\mathbf{F}^{OH}$ is fed into an hourglass-structured hand encoder to produce MANO-related features and regress 2D hand joint coordinates.

3.3. Occlusion-invariant Feature Learning

Hands are frequently occluded when interacting with objects. To achieve robust estimation, the extracted hand feature for the same hand pose should be occlusion-invariant and irrespective of object presence. Given the fact that estimating the bare hand pose is easier than that of the one occluded by a held object, our key insight is to enable the network to simulate the non-occluded hand features from the occluded ones by transferring cross-domain knowledge. Thus, we propose generating plausible de-occluded hand images as pairs for training samples affected by object-caused occlusions, employing an adaptive adjustment strategy for control strength to maximize generation quality. Further, we design multi-level feature enhancement techniques that leverage this paired data to promote comprehensive hand feature learning.

Generative De-occluder. For occluded hand images, our goal is to de-occlude by realistically removing the grasped object while preserving the hand pose to create paired data. Inspired by [33, 34], we utilize ControlNet [58], pre-trained on synthetic hand images, to repaint the hand-object region $\mathbf{M}$ guided by a rendered hand depth map $\mathbf{D}$. Specifically, following the latent diffusion model [42], the original image $\mathbf{X}$ is first projected into latent space as $\mathbf{x}_0$ using a variational auto-encoder [27]. We then follow the standard forward diffusion process as outlined in [21]. In each reverse step $t \in \{T, T-1, ..., 1\}$, to preserve the known background region $(1-\mathbf{M}) \odot \mathbf{X}$, we can alternate the corresponding feature using $(1-\mathbf{m}) \odot \mathbf{x}_t$ as long as maintaining the correct properties of its distribution, as the transition from $\mathbf{x}_t$ to $\mathbf{x}_{t-1}$ depends solely on $\mathbf{x}_t$, *i.e.*,

$$\mathbf{x}_{t-1}^{bg} \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t\mathbf{I})), \tag{4}$$

where $\mathbf{m}$ is downsampled from $\mathbf{M}$ for calculations in latent space, and $\odot$ is the element-wise product. $\mathcal{N}(\cdot)$ denotes the Gaussian distribution. $\bar{\alpha}_t$ denotes the total noise variance at step t, as defined in [44]. For the unknown hand-object region $\mathbf{M} \odot \mathbf{X}$, we perform the reverse diffusion process using the DDIM sampler [44], *i.e.*,

$$\mathbf{x}_{t-1}^{ho} = \text{DDIM}(\epsilon_\theta(\mathbf{x}_t, \mathbf{x}_{mask}, \mathbf{D})), \tag{5}$$

where $\mathbf{x}_{mask}$ is the latent feature masked by $\mathbf{m}$. $\epsilon_\theta(\cdot)$ denotes the denoising model. Thus, the final expression of $\mathbf{x}_{t-1}$ during one reverse step is:

$$\mathbf{x}_{t-1} = \mathbf{m} \odot \mathbf{x}_{t-1}^{ho} + (1-\mathbf{m}) \odot \mathbf{x}_{t-1}^{bg}, \tag{6}$$

which means $\mathbf{x}_{t-1}^{bg}$ is sampled using the known background pixels, while $\mathbf{x}_{t-1}^{ho}$ is sampled from the bare-hand data distribution. They are combined into the new $\mathbf{x}_{t-1}$ using the hand-object mask, ensuring both consistency and realism. After the iterative reverse process, the final denoised vector $\mathbf{x}_0$ is sent to the decoder [27] to recover images from latent features. We show examples of different occlusion conditions along with their de-occluded counterparts in Fig. 3.

Adaptive Control Strength Adjustment. To balance consistency with the condition and realism of the generated hand images, the user often needs to manually adjust the control strength parameter in the generative model. We visualize some examples generated using different control strengths in Fig. 4. In certain cases, such as the top row, too-small control strengths make the generated hand not align well with the depth condition (see (a-b)), while overly-large strengths result in unrealistic appearances (see (d)). Conversely, in other cases like the bottom row, a large control strength is needed to ensure correct hand anatomy (see (h)). Therefore, a fixed control strength cannot be universally applied, while manually setting the value for each case is impractical. To this end, we propose adaptively and automatically adjusting the control strength to enhance generation quality. Specifically, we first define candidate control strengths $\{s_1, s_2, ..., s_n \mid 0 < s_i \leq 1, \forall i\}$ and generate a de-occluded image for each. Then, we employ a pre-trained hand reconstruction model from [9] to estimate the 3D hand poses from the generated images and evaluate the J-PE against the ground truth. The generated sample with the lowest J-PE is incorporated into the training process with the same ground-truth labels as the original sample, *e.g.*, (c) and (h) in Fig. 4 are selected. This approach maximizes the generation quality by selecting a proper control strength that best balances realism and consistency for each case.

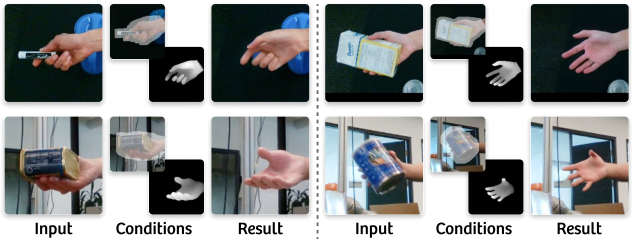


Figure 3. De-occluded examples in various occlusion conditions.

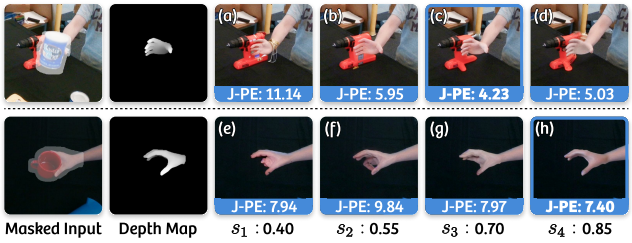


Figure 4. Visualization of our adaptive control strength adjustment.

Multi-level Feature Enhancement. Given the pair of the original image and the corresponding generated image $(\mathbf{X}, \mathbf{X}')$, our goal is to enhance the hand feature representation of $\mathbf{X}$ by leveraging information from $\mathbf{X}'$. To achieve this, we introduce pair-wise feature constraints within a single network in a self-distillation manner. To holistically enhance the capability of recovering occluded information for hand, as shown in Fig. 2, we enhance hand features throughout the hand branch at multiple levels: (i) The initial-level feature $\mathbf{F}_i^h$, extracted from the initial layer of the FPN, captures the low-level information of the hand; (ii) the RoI-level feature $\mathbf{F}_r^h$, output from the FPN after the RoIAlign operation, is utilized for adaptive fusion with the object feature; and (iii) the MANO-level feature $\mathbf{F}_m^h$, extracted before the MANO decoder, serves as the most pertinent feature for regressing the MANO coefficients. Since the MANO-level feature is at a relatively late stage, the occluded and de-occluded counterparts may not always reside in a similar feature space. Inspired by [7], we adopt a multi-head attention block $h(\cdot)$ as the adaptation layer to improve knowledge transfer. Overall, the feature enhancement constraints are formulated as the L1 loss between the features of $\mathbf{X}$ and $\mathbf{X}'$:

$$\begin{aligned} &\mathcal{L}_{init}^{enh} = ||\mathbf{F}_i^h - \mathbf{F}_i^{h'}||_1, \quad \mathcal{L}_{RoI}^{enh} = ||\mathbf{F}_r^h - \mathbf{F}_r^{h'}||_1, \\ &\text{and} \quad \mathcal{L}_{MANO}^{enh} = ||h(\mathbf{F}_m^h) - h(\mathbf{F}_m^{h\,\prime})||_1, \end{aligned} \tag{7}$$

where features with primes belong to the generated image.

Occlusion-aware Case Filtering. During the knowledge transition from the generated samples to the original ones, we observe that the feature learning process may not benefit if the original hand is already non-occluded. In this case, the knowledge gap between de-occluded and occluded hand features disappears, which causes the feature constraints to focus on mitigating the sim-to-real domain gap, making the original features close to the simulated ones, thus yielding suboptimal performance. To address this issue, we filter out non-occluded samples for paired feature enhancement. Specifically, we compute the Intersection over Union (IoU) of the provided amodal hand mask (considering object-caused occlusions) and the rendered full hand mask as the ground-truth occlusion proportion. During training, feature constraints are exclusively applied to pairs whose original occlusion proportion $\hat{O}$ exceeds a pre-defined threshold τ. Furthermore, non-grasping samples are excluded as they do not contribute additional information. Thus, the feature enhancement is conducted only on grasping and occluded samples, *i.e.*,

$$\mathcal{L}_*^{enh} = \mathbb{1}((\hat{O} \geq \tau) \wedge (\hat{G} = 1)) \cdot \mathcal{L}_*^{enh}, \tag{8}$$

where $\tau = 0.1$ in our experiments. $\mathcal{L}_*^{enh}$ represents the feature enhancement constraints at multiple levels.

3.4. Loss Functions

Our training loss includes regular hand and object losses similar to [31], along with our object switcher loss $\mathcal{L}^s$, and feature enhancement constraints $\mathcal{L}_{init}^{enh}$, $\mathcal{L}_{RoI}^{enh}$, and $\mathcal{L}_{MANO}^{enh}$. First, the hand loss is computed as $\mathcal{L}^h = \mathcal{L}^J + \mathcal{L}^V + \mathcal{L}^{MANO}$, where

$$\begin{aligned} \mathcal{L}^J &= ||\mathbf{J}^{2D} - \hat{\mathbf{J}}^{2D}||_2 + ||\mathbf{J}^{3D} - \hat{\mathbf{J}}^{3D}||_2, \\ \mathcal{L}^V &= ||\mathbf{V} - \hat{\mathbf{V}}||_2, \quad \text{and} \quad \mathcal{L}^{MANO} = ||(\theta;\beta) - (\hat{\theta};\hat{\beta})||_2. \end{aligned} \tag{9}$$

Here $\mathbf{J}^{2D}$, $\mathbf{J}^{3D}$, and $\mathbf{V}$ represent the 2D joint, 3D joint, and 3D vertex coordinates, respectively. $(\theta;\beta)$ represent the MANO coefficients. The hat superscript denotes the ground-truth label.

The object loss $\mathcal{L}^o$ supervises the predictions of the 2D location (projected from 3D object keypoints) and their corresponding confidences from image grid proposals [41], *i.e.*,

$$\mathcal{L}^o = \sum_g \sum_{k=1}^{N^o} (||p_{g,k} - \hat{p}_{g,k}||_1 + ||c_{g,k} - \hat{c}_{g,k}||_1), \tag{10}$$

where N^o is the number of keypoints in the 3D bounding box of object mesh. $p_{g,k}$ and $c_{g,k}$ are the pixel location and confidence value at the grid g and control point k, respectively. The hat superscript denotes the ground truth. We compute the object loss only for those grasping images, as they contain the complete object for pose estimation.

Overall, the training loss is as follows,

$$\begin{aligned} \mathcal{L}^{total} = &\underbrace{\mathcal{L}^h + \mathcal{L}^o + \alpha\mathcal{L}^s}_{\text{sample-wise}} \\ &+ \underbrace{\gamma_{init}\mathcal{L}_{init}^{enh} + \gamma_{RoI}\mathcal{L}_{RoI}^{enh} + \gamma_{MANO}\mathcal{L}_{MANO}^{enh}}_{\text{pair-wise}} \end{aligned} \tag{11}$$

where α and γ_* are weights to balance the loss terms.

	Methods	*All Scenes*				*Hand-Only Scene*				*Hand-Object Scene*			
		J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓
HPE	HandOccNet [39]	14.02	6.17	13.55	5.95	13.16	5.31	12.70	5.11	14.58	6.73	14.10	6.49
	MobRecon [9]	15.02	6.71	13.93	5.91	14.43	5.88	13.45	5.15	15.40	7.25	14.24	6.39
	H2ONet [49]	14.78	5.72	14.63	6.19	14.14	4.74	14.00	5.35	15.20	6.35	15.03	6.74
	SimpleHand [64]	14.78	6.30	14.11	6.03	14.63	5.62	13.96	5.38	14.88	6.74	14.21	6.45
HOPE	Liu *et al.* [32]	15.33	6.17	14.79	5.98	15.18	5.48	14.60	5.31	15.43	6.61	14.91	6.40
	Keypoint Trans. [18]	19.16	7.70	18.71	7.96	19.75	7.59	19.26	7.98	18.79	7.77	18.35	7.94
	HFL-Net [31]	14.32	6.08	13.83	5.86	13.61	5.20	13.10	5.01	14.77	6.64	14.29	6.41
Unified	H2ONet† + HFL-Net†	14.12	5.83	13.75	5.83	13.29	**4.70**	13.08	5.06	14.66	6.55	14.18	6.33
	H2ONet‡ + HFL-Net‡	14.54	5.90	14.19	6.00	14.14	4.76	14.00	5.35	14.79	6.63	14.31	6.41
	HandOccNet† + HFL-Net†	14.34	6.06	13.86	5.85	13.85	5.28	13.35	5.08	14.66	6.56	14.18	6.34
	HandOccNet‡ + HFL-Net‡	14.52	6.15	14.02	5.93	14.09	5.37	13.58	5.17	14.79	6.65	14.31	6.42
	UniHOPE (ours)	**13.03**	**5.59**	**12.59**	**5.40**	**12.59**	4.83	**12.12**	**4.66**	**13.31**	**6.08**	**12.89**	**5.87**

Table 2. Hand-pose estimation results on DexYCB. †: pre-trained in the original setting. ‡: re-trained in the unified setting. The best and second-best are marked in **bold** and underlined. Our UniHOPE attains leading performance for almost all metrics in all scenarios.

Methods	gelatin_box	bleach_cleanser	wood_block	average ↑
Liu *et al.* [32]	26.31	25.07	68.56	38.89
Keypoint Trans. [18]	0.00	1.31	32.61	10.47
HFL-Net [31]	25.88	32.08	70.16	41.59
H2ONet† + HFL-Net†	29.26	30.71	64.84	40.69
H2ONet‡ + HFL-Net‡	26.12	29.33	69.40	40.51
HandOccNet† + HFL-Net†	**29.30**	30.57	64.94	40.69
HandOccNet‡ + HFL-Net‡	26.22	29.51	69.40	40.61
UniHOPE (ours)	26.23	**32.32**	**74.29**	**43.06**

Table 3. Unseen object-pose estimation results on DexYCB.

	Methods	J-PE ↓	J-AUC ↑	V-PE ↓	V-AUC ↑	F@5 ↑	F@15 ↑
HPE	HandOccNet [39]	28.94	47.96	28.10	49.23	23.34	67.86
	MobRecon [9]	29.61	48.10	28.65	49.59	23.38	67.72
	H2ONet [49]	30.46	47.09	29.55	48.31	22.10	66.04
	SimpleHand [64]	29.01	47.58	28.09	49.00	21.93	65.90
HOPE	Liu *et al.* [32]	29.54	47.55	28.66	48.83	21.82	67.34
	Keypoint Trans. [18]	41.04	34.47	39.64	36.02	17.13	56.07
	HFL-Net [31]	28.45	50.34	27.55	51.57	24.31	69.88
Unified	H2ONet† + HFL-Net†	31.27	47.70	30.31	48.86	22.91	67.20
	H2ONet‡ + HFL-Net‡	28.49	50.45	27.59	51.67	24.33	69.86
	HandOccNet† + HFL-Net†	30.96	47.85	30.01	49.02	23.04	67.47
	HandOccNet‡ + HFL-Net‡	28.33	50.45	27.44	51.67	24.39	69.99
	UniHOPE (ours)	**26.23**	**52.26**	**25.41**	**53.52**	**24.64**	**70.77**

Table 4. Hand-pose estimation results (*Root-relative*) on HO3D.

4. Experiments

4.1. Experimental Settings

Datasets. In our unified setting, we organize the original dataset into hand-only and hand-object scenes based on the object grasping status. We conduct experiments on the following commonly-used datasets: (i) **DexYCB** [6]: we use the more challenging "S3" split (train/test: 376,374/76,360 samples) with unseen grasped objects in the test set (train/test: 15/3 objects). We report performance on the entire dataset (all scenes) as well as separately for hand-only and hand-object scenes; (ii) **HO3D** [17] (version 2, train/test: 66,034/11,524 samples): results are submitted to the online server as the ground-truth 3D hand annotations are not publicly accessible, hence results for each scene are unavailable. Further, to evaluate the generalization ability, we perform cross-dataset validation on the test set of (iii) **FreiHAND** [65] (train/test: 130,240/3,960 samples), which mainly consists of bare-hand images and lacks object annotations for scene division. Consequently, results for each scene are also unavailable. More details and results on other data splits of DexYCB are provided in the Supp.

Evaluation Metrics. We evaluate hand pose estimation using commonly-used metrics, as in [9, 39, 49]: (i) J/V-PE denotes the mean per joint/vertex position error (also known as MPJPE/MPVPE) in mm measured by Euclidean distance between estimated and ground-truth 3D hand joint/vertex coordinates; (ii) J/V-AUC calculates the area under the curve (in percentage) of the percentage of correct keypoints (PCK) across different error thresholds for joint/vertex; and (iii) F@5/F@15 is the harmonic mean of recall and precision (in percentage) between estimated and ground-truth 3D hand vertices under 5mm/15mm thresholds. We report these metrics both before (*i.e.*, root-relative) and after Procrustes Alignment (PA), which aligns the estimation with ground truths by global rotation, translation, and scale adjustment.

For object pose evaluation, we measure the average 3D distance (ADD) of the grasped object. Specifically, we report ADD-0.5D, the percentage of objects whose ADD is within 50% of the object diameter as in [23] considering the challenge of unseen-object pose estimation in DexYCB.

Implementation Details. We train UniHOPE on eight NVidia RTX 2080Ti GPUs using a batch size of 64 and the Adam optimizer [28] with an initial learning rate of 1e-4 (decay by 0.7 every 10 epochs). Input images are resized to 128×128 and augmented with random scaling, rotating, translating, and color jittering. To stabilize training, we first train the network with both original and generated images for 30 epochs, then incorporate the feature enhancement constraints for another 40 epochs. Please refer to our Supp. for more details.

4.2. Comparison with SOTA Methods

We compare our UniHOPE with previous SOTA HPE and HOPE methods [9, 18, 32, 39, 64] in our unified setting (trained using their officially-released code). To support

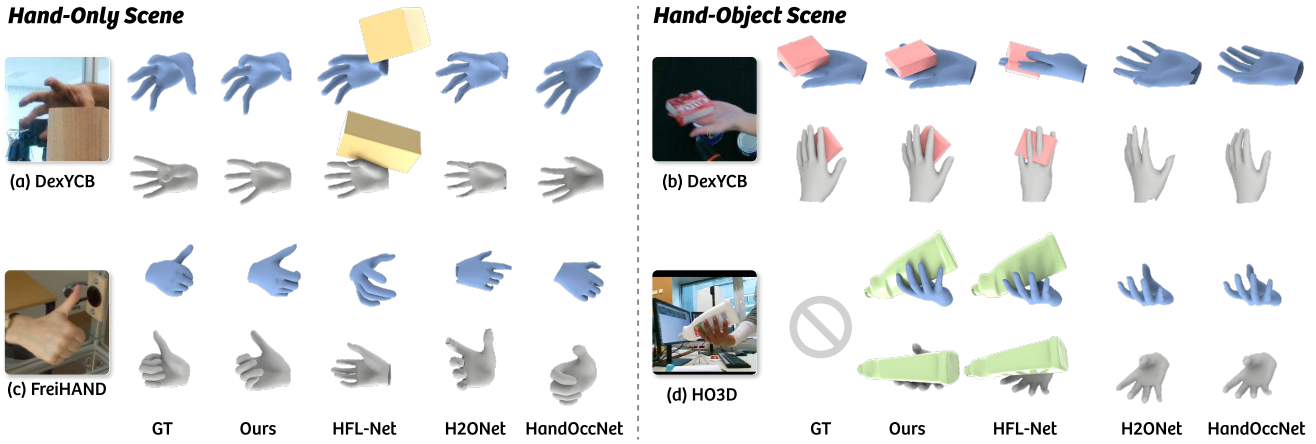


Figure 5. Qualitative comparison between our method and SOTA HPE/HOPE methods on hand-only/hand-object scenarios across different datasets. The first and second rows in each example denote the original view and another view, respectively, for better comparison.

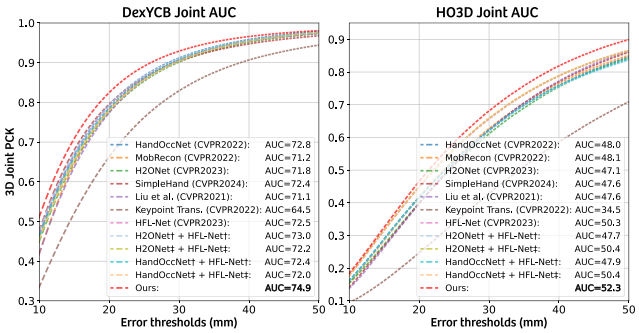


Figure 6. The joint AUC comparison under different thresholds. UniHOPE achieves better performance than others, consistently.

the unified task, one straightforward solution is to use a classifier to determine if the hand is grasping any object, then employ an existing HPE or HOPE method accordingly. We denote the combination as A + B, where A is the SOTA HPE method H2ONet [49] or HandOccNet [39], and B is the SOTA HOPE method, HFL-Net [31].

Evaluation on DexYCB. We present quantitative comparisons of hand pose estimation on DexYCB in Tab. 2. Our UniHOPE achieves the best performance overall, showing its effectiveness in general hand-object interactions across various scenarios. The root-relative 3D joint PCK/AUC comparison under different thresholds is shown in Fig. 6, further confirming the comprehensive performance of our approach.

For object pose estimation accuracy, we conduct comparisons using per-instance and average ADD-0.5D, as shown in Tab. 3. Compared with SOTA HOPE methods, our method achieves the highest average score on the test set with unseen objects, highlighting its superiority in estimating object pose. Qualitative comparisons with SOTA methods [31, 39, 49] are illustrated in Fig. 5. In the hand-only scenario (see Fig. 5 (a)), where the wood block has not been grasped yet, SOTA HOPE method [31] inevitably produces an extra object pose due to its inflexibility; in contrast, our method does not, thanks to our object switcher's >95% grasping status prediction accuracy. Moreover, our method yields more plausible hand poses. In the hand-object scenario (see Fig. 5 (b)), our method produces high-quality hand-object poses, whereas previous methods fail when the hand experiences moderate occlusion, indicating our robustness against such challenges.

Evaluation on HO3D. We conduct the same experiment on HO3D [17]. Tab. 4 shows the root-relative quantitative comparison. Our method achieves top performance across all metrics, demonstrating its effectiveness and robustness. The joint PCK/AUC curve in Fig. 6 also confirms the consistent best results of our approach. In addition, qualitative comparisons in Fig. 5 (d) clearly illustrate the superiority of our method in estimating 3D hand and object pose under partial object-caused occlusion. More quantitative results are available in the Supp.

Evaluation on FreiHAND. To assess generalization ability, we perform cross-dataset validation by transferring models trained on DexYCB to the FreiHAND test set. As reported in Tab. 5, our method outperforms all SOTA HPE/HOPE methods, particularly by a substantial margin in root-relative metrics. The qualitative comparison in Fig. 5 (c) also shows that UniHOPE produces more accurate hand poses in challenging cases, indicating improved generalization to unseen bare-hand scenes.

Evaluation under Different Levels of Occlusion. To showcase the robustness of our method against object-caused occlusion, we partition the DexYCB test set into different occlusion levels based on the ground-truth hand-object occlusion proportion (as detailed in Sec. 3.3) and provide quantitative comparisons in Tab. 6. Our UniHOPE exhibits the best hand pose estimation performance across all occlusion levels, underscoring the efficacy of our feature enhancement techniques. Note that test samples where the hand being absent from the image region are excluded.

	Methods	*Root-relative*						*Procrustes Alignment*					
		J-PE ↓	J-AUC ↑	V-PE ↓	V-AUC ↑	F@5 ↑	F@15 ↑	J-PE ↓	J-AUC ↑	V-PE ↓	V-AUC ↑	F@5 ↑	F@15 ↑
HPE	HandOccNet [39]	58.03	29.07	56.06	29.40	14.89	47.75	14.52	71.34	14.09	72.07	40.02	88.08
HPE	MobRecon [9]	71.18	22.21	68.50	22.34	10.41	36.67	17.74	65.16	17.32	65.90	32.95	81.35
HPE	H2ONet [49]	79.14	19.88	76.29	19.38	9.69	36.71	14.56	71.21	14.25	71.69	38.73	87.69
HPE	SimpleHand [64]	60.82	26.09	58.85	26.34	12.84	43.53	15.79	68.90	15.37	69.61	37.03	85.40
HOPE	Liu *et al*. [32]	59.40	28.84	57.39	29.04	14.92	47.76	14.00	72.30	13.59	73.02	41.53	89.11
HOPE	Keypoint Trans. [18]	96.27	13.60	93.39	12.37	6.99	27.82	16.97	66.66	17.03	66.38	34.31	83.44
HOPE	HFL-Net [31]	58.02	29.94	56.08	30.20	15.47	48.83	14.29	71.75	13.85	72.52	40.52	88.63
Unified	H2ONet† + HFL-Net†	68.88	24.27	66.49	24.21	12.33	42.38	14.47	71.36	14.10	71.99	39.48	88.14
Unified	H2ONet‡ + HFL-Net‡	68.25	24.49	65.90	24.40	12.30	42.51	14.42	71.47	14.06	72.09	39.54	88.18
Unified	HandOccNet† + HFL-Net†	81.79	21.33	78.91	20.56	10.89	36.68	15.40	69.65	14.85	70.58	38.30	86.94
Unified	HandOccNet‡ + HFL-Net‡	80.74	20.30	78.00	19.44	10.52	36.38	16.15	68.29	15.58	69.26	36.92	85.93
Unified	UniHOPE (ours)	**50.97**	**34.31**	**49.21**	**34.94**	**17.83**	**53.46**	**13.53**	**73.24**	**13.14**	**73.92**	**43.23**	**89.55**

Table 5. Cross-dataset validation of hand-pose estimation on FreiHAND.

Methods	*Occlusion (25%-50%)*				*Occlusion (50%-75%)*				*Occlusion (75%-100%)*			
	J-PE↓	PA-J-PE↓	V-PE ↓	PA-V-PE ↓	J-PE↓	PA-J-PE↓	V-PE ↓	PA-V-PE ↓	J-PE↓	PA-J-PE↓	V-PE ↓	PA-V-PE ↓
HandOccNet [39]	16.40	7.08	15.85	6.83	18.22	7.60	17.67	7.33	28.15	8.71	27.20	8.40
MobRecon [9]	16.67	7.61	15.60	6.77	20.04	8.17	18.80	7.48	31.46	9.64	29.97	9.23
H2ONet [49]	17.07	6.76	16.78	7.09	19.41	7.32	19.07	7.58	31.07	8.82	30.11	8.94
SimpleHand [64]	16.43	7.05	15.78	6.75	19.33	7.55	18.38	7.39	38.52	10.57	36.85	10.40
Liu *et al*. [32]	16.98	6.92	16.43	6.71	19.72	7.11	19.14	6.91	33.80	8.99	32.64	8.71
Keypoint Trans. [18]	20.95	8.15	20.41	8.31	24.45	8.61	23.88	8.76	38.29	11.21	37.39	11.75
HFL-Net [31]	16.33	7.00	15.81	6.77	18.66	7.33	18.11	7.11	28.95	8.80	27.94	8.53
H2ONet† + HFL-Net†	16.07	6.84	15.57	6.67	19.39	7.40	18.82	7.22	30.32	8.43	29.27	8.28
H2ONet‡ + HFL-Net‡	16.43	6.93	15.94	6.79	18.76	7.29	18.25	7.14	31.02	8.55	29.91	8.45
HandOccNet† + HFL-Net†	16.04	6.89	15.53	6.65	19.22	7.43	18.64	7.20	29.57	8.58	28.53	8.31
HandOccNet‡ + HFL-Net‡	16.41	7.00	15.88	6.77	18.64	7.33	18.09	7.10	29.22	8.76	28.20	8.47
UniHOPE (ours)	**14.59**	**6.39**	**14.13**	**6.17**	**16.27**	**6.51**	**15.78**	**6.29**	**26.42**	**7.64**	**25.51**	**7.40**

Table 6. Comparison with SOTA methods across different object-caused occlusion levels on DexYCB.

	Models	*Root-relative*		*Procrustes Align.*	
		J-PE ↓	V-PE ↓	J-PE ↓	V-PE ↓
(a)	Baseline	14.09	13.61	5.95	5.75
(b)	w/ Grasp-aware Feature Fusion	13.84	13.37	5.79	5.58
(c)	w/ Generative De-occluder	13.38	12.92	5.71	5.52
(d)	+ Image Feature Enhancement	13.23	12.79	5.64	5.44
(e)	+ RoI Feature Enhancement	13.18	12.73	5.64	5.45
(f)	+ MANO Feature Enhancement	13.12	12.67	5.63	5.43
(g)	+ Occlusion-aware Case Filtering	**13.03**	**12.59**	**5.59**	**5.40**

Table 7. Ablation study on major designs in UniHOPE.

4.3. Ablation Studies

We perform ablation studies on DexYCB to evaluate the effectiveness of our designs, as shown in Tab. 7. HFL-Net [31] with our object switcher serves as the simplest baseline.

Grasp-aware Feature Fusion. We first analyze the impact of the grasp-aware feature fusion module. Comparison of Rows (a-b) shows performance boosts across all metrics, indicating that integrating irrelevant object features affects hand pose estimation and our design alleviates this issue.

Generative De-occluder. Next, we assess the effects of de-occluded hand images with in-distribution hand poses, as they provide extra information. Comparing Rows (b-c), the notable improvement upon incorporating paired data indicates the effectiveness of synthetic samples. More details on our adaptive control strength adjustment are in the Supp.

Occlusion-invariant Feature Learning. Further, we show the individual effects of feature enhancement at various levels. Comparing Row (c) with (d-f), we note a progressive improvement in root-relative metrics at each level, showing the efficacy of knowledge transferring from de-occluded hands. Finally, Row (g) reveals that the effects are maximally realized through our occlusion-aware case filtering.

5. Conclusion

We introduce UniHOPE, the first unified approach for hand-only and hand-object pose estimation, motivated by the inability of existing methods to handle both scenes. Our technical innovations are twofold: first, to enable flexibility in switching between different scenes, we incorporate an object switcher to control object-pose estimation and design a grasping-aware feature fusion module to selectively capture effective object features; second, to promote robustness against object-caused occlusion, we propose multi-level feature enhancement to learn occlusion-invariant hand features from generated realistic de-occluded hand images. Experimental results on three common benchmarks manifest the SOTA performance of UniHOPE.

Acknowledgments This work is supported in part by the Research Grants Council of the Hong Kong Special Administrative Region, China, under Project T45-401/22-N and by the Hong Kong Innovation and Technology Fund, under Project MHP/086/21.

References

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

[4] 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

[5] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *ICCV*, pages 12417–12426, 2021. 1

[6] 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. 3, 6

[7] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *NeurIPS*, 30, 2017. 5

[8] 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

[9] 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. 1, 2, 4, 6, 8

[10] 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

[11] Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. AlignSDF: Pose-aligned signed distance fields for hand-object reconstruction. In *ECCV*, pages 231–248. Springer, 2022. 2

[12] Zerui Chen, Shizhe Chen, Cordelia Schmid, and Ivan Laptev. gSDF: Geometry-driven signed distance functions for 3D hand-object reconstruction. In *CVPR*, pages 12890–12900, 2023. 2

[13] Hongsuk Choi, Nikhil Chavan-Dafle, Jiacheng Yuan, Volkan Isler, and Hyunsoo Park. HandNeRF: Learning to reconstruct hand-object interaction scene from a single RGB image. In *ICRA*, pages 13940–13946. IEEE, 2024. 2

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

[15] 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. 2

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[17] 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. 3, 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*, pages 11090–11100, 2022. 1, 2, 6, 8

[19] 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

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, pages 2961–2969, 2017. 4

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 2, 4

[22] Hezhen Hu, Weilun Wang, Wengang Zhou, and Houqiang Li. Hand-object interaction image generation. *NeurIPS*, 35: 23805–23817, 2022. 3

[23] Yinlin Hu, Pascal Fua, and Mathieu Salzmann. Perspective flow aggregation for data-limited 6D object pose estimation. In *ECCV*, pages 89–106. Springer, 2022. 6

[24] Di Huang, Xiaopeng Ji, Xingyi He, Jiaming Sun, Tong He, Qing Shuai, Wanli Ouyang, and Xiaowei Zhou. Reconstructing hand-held objects from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 2

[25] Lin Huang, Chung-Ching Lin, Kevin Lin, Lin Liang, Lijuan Wang, Junsong Yuan, and Zicheng Liu. Neural voting field for camera-space 3D hand pose estimation. In *CVPR*, pages 8969–8978, 2023. 1, 2

[26] 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

[27] Diederik P. Kingma. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

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

[29] 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. 2

[30] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 3

[31] Zhifeng Lin, Changxing Ding, Huan Yao, Zengsheng Kuang, and Shaoli Huang. Harmonious feature learning for interactive hand-object pose estimation. In *CVPR*, pages 12989–12998, 2023. 1, 2, 4, 5, 6, 7, 8

[32] 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. 1, 2, 4, 6, 8

[33] Wenquan Lu, Yufei Xu, Jing Zhang, Chaoyue Wang, and Dacheng Tao. HandRefiner: Refining malformed hands in generated images by diffusion-based conditional inpainting. In *ACM MM*, 2024. 3, 4

[34] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, pages 11461–11471, 2022. 4

[35] Matthew Matl. Pyrender. https://github.com/mmatl/pyrender, 2019. 2

[36] 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. 1, 2

[37] 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

[38] Supreeth Narasimhaswamy, Uttaran Bhattacharya, Xiang Chen, Ishita Dasgupta, Saayan Mitra, and Minh Hoai. HanDiffuser: Text-to-image generation with realistic hand appearances. In *CVPR*, pages 2468–2479, 2024. 3

[39] 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. 1, 2, 6, 7, 8

[40] Junho Park, Kyeongbo Kong, and Suk-Ju Kang. AttentionHand: Text-driven controllable hand image generation for 3D hand reconstruction in the wild. *arXiv preprint arXiv:2407.18034*, 2024. 3

[41] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 5

[42] 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, 4

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

[44] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 4

[45] 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, 4

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, pages 6000–6010, 2017. 4

[47] Rong Wang, Wei Mao, and Hongdong Li. Interacting hand-object pose estimation via dense mutual attention. In *WACV*, pages 5735–5745, 2023. 1

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

[49] 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. 1, 2, 6, 7, 8

[50] Hao Xu, Haipeng Li, Yinqiao Wang, Shuaicheng Liu, and Chi-Wing Fu. HandBooster: Boosting 3D hand-mesh reconstruction by conditional synthesis and sampling of hand-object interactions. In *CVPR*, pages 10159–10169, 2024. 3

[51] 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

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

[53] 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. 2

[54] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3D reconstruction of generic objects in hands. In *CVPR*, pages 3895–3905, 2022. 2

[55] 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. 3

[56] Yufei Ye, Abhinav Gupta, Kris Kitani, and Shubham Tulsiani. G-HOP: Generative hand-object prior for interaction reconstruction and grasp synthesis. In *CVPR*, pages 1911–1920, 2024. 2

[57] 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

[58] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 4

[59] Mengqi Zhang, Yang Fu, Zheng Ding, Sifei Liu, Zhuowen Tu, and Xiaolong Wang. HOIDiffusion: Generating realistic 3D hand-object interaction data. In *CVPR*, pages 8521–8531, 2024. 3

[60] 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

[61] 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.

[62] 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.

[63] 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

[64] Zhishan Zhou, Shihao Zhou, Zhi Lv, Minqiang Zou, Yao Tang, and Jiajun Liang. A simple baseline for efficient hand mesh reconstruction. In *CVPR*, pages 1367–1376, 2024. 1, 2, 6, 8

[65] 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, 6

arXiv:2503.13303v1 [cs.CV] 17 Mar 2025

UniHOPE: A Unified Approach for Hand-Only and Hand-Object Pose Estimation

Supplementary Material

In this supplementary material, we provide more qualitative and quantitative results to show the capabilities and robustness of UniHOPE (Sec. A). In Sec. B, we present the implementation details and in Sec. C, we discuss the limitations and future work.

A. More Experimental Results

A.1. Qualitative Results

First of all, we present Figs. A to D, which show that UniHOPE is able to handle both hand-only scenario (left columns) and hand-object scenario (right columns).

Comparison with SOTA Methods. Next, we provide more qualitative comparisons on the DexYCB (Fig. E), HO3D (Fig. F), and FreiHAND datasets (Fig. G).

More De-occluded Examples. Furthermore, we present more de-occluded samples in Fig. H.

A.2. Quantitative Results

Additional Results of Tab. 1. The additional metrics of Tab. 1 in the main paper are provided in Tab. A. Both the metrics before & after PA show an overall performance degeneration of existing HPE/HOPE models when transferring to apply to the other scenario or testing in the original scenario even after re-training on both scenes.

Comparison on Other Splits of DexYCB. We provide the quantitative results of hand pose estimation on the default "S0" split (same distribution for the training and test set) and "S1" split with unseen subjects (train/test: 7/2 subjects) of DexYCB in Tab. B and Tab. C, respectively. Our method achieves the best overall performance, especially in root-relative metrics.

Comparison on HO3D. The remaining hand metrics on HO3D are reported in Tab. D. Though HFL-Net [9] and the combination of H2ONet + HFL-Net achieve better PA results, our method outperforms them by a large margin in the metrics after scale-translation only alignment [4], which takes both the global rotation and hand shape into consideration. We emphasize the importance of global rotation, since it better reflects the visualization quality, as indicated by the qualitative comparison results shown in Fig. F.

A.3. Detailed Analysis on Performance

In this work, we explore a new setting to address HPE and HOPE at once. Applying prior SOTA of HPE/HOPE is suboptimal, even re-trained on all scenarios, as they lack specific designs. For hand-only scenes, HOPE methods are affected by irrelevant object features, even no object is grasped, yet HPE methods may fail for unseen hand poses. For hand-object scenes, HOPE methods lack effective designs to handle severe occlusions, while HPE methods do not utilize object information to enhance performance. Our approach works better in each scene type. As Fig. I shows: (a) when the hand reaches out to grasp an object, our grasp-aware feature fusion reduces the adverse impact of non-grasped object; (b) for unseen hand poses from FreiHAND, our generated de-occluded images introduce richer hand poses to boost performance; (c) our multi-level feature enhancement improves robustness under severe object occlusions; and (d) when grasping objects, our method surpasses HPE methods by leveraging object information. These observations are consistent with the quantitative performance in Tab. 2, 5, 6 in the main paper.

A.4. Additional Ablation Studies

To be consistent with the main paper, we conduct all the ablation studies presented below on DexYCB.

Additional Results of Tab. 7. Since the RHD [22] and Static Gestures Dataset [1] are utilized to fine-tune the ControlNet [12], we also conduct an ablation study of pre-training on these synthetic datasets before training on DexYCB, using a network structure identical to our baseline model with the grasp-aware feature fusion module (Row (b) of Tab. 7 in the main paper). As shown in Tab. F, directly incorporating synthetic datasets into training leads to a minor improvement, indicating the limitation caused by the domain gap between the synthetic and real-world images. Conversely, our occlusion-invariant feature learning strategy substantially enhances the model performance through the foundational data prior provided by ControlNet [20] and the multi-level feature enhancement.

Ablation on Adaptive Control Strength Adjustment. Control strength (ranging from 0 to 1) is imposed on the connections between the ControlNet and Stable Diffusion, controlling the extent to which the output is consistent with the control signal. We propose to adaptively adjust its value with MobRecon [3] pre-trained on DexYCB to avoid tedious manual tuning. The default control strength employed in [12] is 0.55. In our work, we empirically select the candidate control strengths from $\{0.25, 0.4, 0.55, 0.7, 0.85, 1.0\}$, with a similar number of candidates as in [12].



Figure A. UniHOPE is able to handle both hand-only (left column) and hand-object scenarios (right column). Here, we show more qualitative results on DexYCB. For each example, the estimation results are rendered from the original (view 1) and another view (view 2) for clear visualization.



Figure B. More qualitative results of UniHOPE on DexYCB.



Figure C. More qualitative results of UniHOPE across hand-only (left column) and hand-object scenarios (right column) on HO3D.



Figure D. More qualitative results of UniHOPE on HO3D.

HPE	*Hand-Only Scene*				*Hand-Only → Hand-Object Scene*				*All → Hand-Only Scene*				*All → Hand-Object Scene*			
	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓
[14]	12.98	5.21	12.52	5.02	19.60 (-6.62)	7.71 (-2.50)	18.95 (-6.43)	7.42 (-2.40)	13.16 (-0.18)	5.31 (-0.10)	12.70 (-0.18)	5.11 (-0.09)	14.58	6.73	14.10	6.49
[18]	13.34	4.69	13.13	5.05	21.98 (-8.64)	7.13 (-2.44)	21.42 (-8.30)	7.27 (-2.22)	14.14 (-0.80)	4.74 (-0.05)	14.00 (-0.87)	5.35 (-0.30)	15.20	6.35	15.03	6.74
[21]	14.05	5.55	13.51	5.31	18.37 (-4.32)	7.42 (-1.87)	17.54 (-4.03)	6.91 (-1.60)	14.63 (-0.58)	5.62 (-0.07)	13.96 (-0.45)	5.38 (-0.07)	14.88	6.74	14.21	6.45
HOPE	*Hand-Object Scene*				*Hand-Object → Hand-Only Scene*				*All → Hand-Object Scene*				*All → Hand-Only Scene*			
	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓
[5]	17.99	7.68	17.57	7.88	25.10 (-7.11)	7.62 (+0.06)	24.40 (-6.83)	7.88 (-0.00)	18.79 (-1.00)	7.77 (-0.09)	18.35 (-0.78)	7.94 (-0.06)	19.75	7.59	19.26	7.98
[9]	14.61	6.56	14.13	6.33	19.39 (-4.78)	5.96 (+0.60)	18.61 (-4.48)	5.75 (+0.58)	14.77 (-0.16)	6.64 (-0.08)	14.29 (-0.16)	6.41 (-0.08)	13.61	5.20	13.10	5.01

Table A. Full metrics of Tab.1 in the main paper.

	Methods	*All Scenes*				*Hand-Only Scene*				*Hand-Object Scene*			
		J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓
HPE	HandOccNet [14]	13.04	5.85	12.61	5.65	13.42	5.39	12.95	5.20	12.79	6.15	12.39	5.95
	MobRecon [3]	14.34	6.50	13.40	5.74	14.57	5.91	13.74	5.29	14.18	6.88	13.19	6.03
	H2ONet [18]	13.89	**5.38**	13.56	5.52	14.10	**4.84**	13.75	5.02	13.76	**5.73**	13.43	5.84
	SimpleHand [21]	13.66	6.02	13.14	5.78	14.48	5.67	13.95	5.46	13.13	6.24	12.62	5.99
HOPE	Liu *et al*. [11]	14.06	5.75	13.57	5.58	14.87	5.47	14.33	5.30	13.53	5.93	13.08	5.75
	Keypoint Trans. [5]	16.61	6.84	16.21	7.05	18.50	7.03	18.00	7.32	15.39	6.71	15.05	6.88
	HFL-Net [9]	13.02	5.58	12.58	5.39	13.41	5.19	12.92	5.00	12.77	5.84	12.35	**5.64**
Unified	H2ONet† + HFL-Net†	13.08	5.47	12.71	5.43	13.81	4.85	13.50	5.06	12.61	5.87	12.20	5.68
	H2ONet‡ + HFL-Net‡	13.30	5.45	12.91	5.40	14.09	4.85	13.74	5.02	12.79	5.83	12.37	**5.64**
	HandOccNet† + HFL-Net†	13.32	5.73	12.87	5.54	14.40	5.50	13.89	5.30	12.63	5.89	12.22	5.69
	HandOccNet‡ + HFL-Net‡	13.43	5.71	12.97	5.51	14.41	5.49	13.90	5.30	12.80	5.85	12.38	5.65
	UniHOPE (ours)	**12.59**	5.54	**12.17**	**5.36**	**12.84**	5.02	**12.38**	**4.85**	**12.42**	5.88	**12.03**	5.69

Table B. Quantitative comparison on DexYCB "S0" split.

	Methods	*All Scenes*				*Hand-Only scene*				*Hand-Object Scene*			
		J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓	J-PE ↓	PA-J-PE ↓	V-PE ↓	PA-V-PE ↓
HPE	HandOccNet [14]	18.33	6.95	17.70	6.71	19.70	6.01	18.95	5.81	17.57	7.47	17.02	7.21
	MobRecon [3]	18.62	7.18	17.73	6.61	19.36	6.27	18.42	5.75	18.21	7.68	17.36	7.09
	H2ONet [18]	18.40	6.40	17.90	6.57	18.92	**5.44**	18.36	5.70	18.11	6.93	17.64	7.05
	SimpleHand [21]	17.38	6.82	16.81	6.73	18.86	6.02	18.14	5.92	16.57	7.26	16.08	7.17
HOPE	Liu *et al*. [11]	17.82	6.46	17.19	6.25	19.12	5.89	18.36	5.69	17.10	**6.77**	16.54	**6.55**
	Keypoint Trans. [5]	21.61	8.15	21.18	8.36	22.84	7.32	22.24	7.59	20.93	8.61	20.60	8.79
	HFL-Net [9]	17.77	6.58	17.16	6.36	18.42	5.72	17.72	5.52	17.41	7.06	16.86	6.82
Unified	H2ONet† + HFL-Net†	17.49	**6.36**	16.94	6.25	19.24	5.50	18.60	5.59	16.54	6.83	16.02	6.61
	H2ONet‡ + HFL-Net‡	17.96	6.48	17.41	6.42	18.92	5.45	18.35	5.69	17.44	7.05	16.89	6.82
	HandOccNet† + HFL-Net†	17.84	6.53	17.22	6.31	20.18	5.95	19.39	5.75	16.55	6.85	16.03	6.62
	HandOccNet‡ + HFL-Net‡	18.63	6.73	17.99	6.50	20.72	6.11	19.91	5.91	17.48	7.06	16.93	6.83
	UniHOPE (ours)	**16.84**	6.42	**16.25**	**6.20**	**17.80**	5.50	**17.11**	**5.30**	**16.31**	6.93	**15.79**	6.70

Table C. Quantitative comparison on DexYCB "S1" split.

To assess the effectiveness of our adaptive control strength adjustment, we compare our model (Row (c) of Tab. 7 in the main paper) with the ones trained with generated samples under fixed control strengths without incorporating the feature enhancement constraints. As shown in Tab. E, our adaptive strategy achieves the best performance in hand pose estimation compared to several control strengths. The samples generated under all candidate control strengths are provided in Fig. J, showing the need to adaptively select control strength for different cases.

Effects of Hyperparameters. The default value of hyperparameter α is empirically set to 10 in Eq. (11) of the main paper. This is to ensure a prediction accuracy over 95%. For the hyperparameters controlling the feature enhancement at three different levels, we evaluate their effects on the hand pose estimation performance in Tab. G. Since the MANO-level feature is a late-stage feature employed to directly regress the final hand pose, an adaption layer is deployed to improve the knowledge transfer. We set a larger value for γ_{MANO} to aim to strongly enforce this feature adaptation process. In our experiments, the values for γ_{init}, γ_{RoI}, and γ_{MANO} are set to 0.1, 0.1, and 0.5, respectively.

A.5. Computational Cost and Efficiency

The training time of our model is 3 days for DexYCB (376k samples) and 12 hours for HO3D (66k samples), respectively, on eight NVidia RTX 2080Ti GPUs.

Tab. H reports the inference speed (FPS, tested on a sin-

	Methods	*Procrustes Alignment*						*Scale-Translation Aligned*	
		J-PE ↓	J-AUC ↑	V-PE ↓	V-AUC ↑	F@5 ↑	F@15 ↑	J-PE ↓	J-AUC ↑
HPE	HandOccNet [14]	10.26	7.95	10.21	79.61	50.61	94.47	28.18	49.28
	MobRecon [3]	10.47	79.14	10.76	78.54	47.57	93.59	29.36	49.36
	H2ONet [18]	9.52	80.97	9.60	80.81	52.62	95.09	29.67	48.53
	SimpleHand [21]	11.28	77.66	11.58	77.05	45.78	91.74	28.41	49.32
HOPE	Liu *et al*. [11]	9.46	81.12	9.39	81.25	54.93	95.64	28.44	49.79
	Keypoint Trans. [5]	12.00	76.24	12.18	75.83	44.71	91.60	40.00	36.36
	HFL-Net [9]	9.01	82.02	8.92	82.18	57.01	96.19	27.97	51.33
Unified	H2ONet† + HFL-Net†	9.49	81.04	9.43	81.16	54.54	95.54	30.60	48.93
	H2ONet‡ + HFL-Net‡	**8.97**	**82.10**	**8.88**	**82.26**	**57.08**	**96.22**	28.00	51.44
	HandOccNet† + HFL-Net†	9.56	80.89	9.50	81.02	54.23	95.47	30.29	49.09
	HandOccNet‡ + HFL-Net‡	9.05	81.94	8.96	82.10	56.79	96.14	27.83	51.45
	UniHOPE (ours)	9.60	80.82	9.45	81.12	52.57	95.68	**25.53**	**53.70**

Table D. Quantitative comparison (*Procrustes Alignment & Scale-Translation Aligned*) on HO3D.

Control Strength Selection	*Root-relative*		*Procrustes Align.*	
	J-PE ↓	V-PE ↓	J-PE ↓	V-PE ↓
s = 0.4	13.76	13.30	5.85	5.65
s = 0.55	13.51	13.06	5.78	5.57
s = 0.7	13.43	12.98	5.75	5.55
Adaptive Adjustment (ours)	**13.38**	**12.92**	**5.71**	**5.52**

Table E. Quantitative results of our adaptive control strength adjustment *vs*. fixed control strengths.

Models	*Root-relative*		*Procrustes Align.*	
	J-PE ↓	V-PE ↓	J-PE ↓	V-PE ↓
Baseline w/ Grasp-aware Feature Fusion	13.84	13.37	5.79	5.58
w/ RHD [22] & Static Gestures [1]	13.79	13.32	5.73	5.53
Ours	**13.03**	**12.59**	**5.59**	**5.40**

Table F. Comparison with directly training with synthetic datasets used by [12].

γ_{init}/ γ_{RoI} / γ_{MANO}	*Root-relative*		*Procrustes Align.*	
	J-PE ↓	V-PE ↓	J-PE ↓	V-PE ↓
0.001 / 0.001 / 0.005	13.17	12.72	5.61	5.41
0.01 / 0.01 / 0.05	13.13	12.69	5.62	5.42
0.1 / 0.1 / 0.5 (ours)	**13.03**	**12.59**	**5.59**	**5.40**
1.0 / 1.0 / 5.0	13.15	12.70	5.70	5.50
10.0 / 10.0 / 50.0	14.13	13.65	6.08	5.87

Table G. Effects of various hyperparameters of the multi-level feature constraints.

gle NVidia RTX 2080Ti GPU), FLOPs, and number of parameters of various models. Thanks to the lightweight object switcher in UniHOPE, UniHOPE has similar inference efficiency and model complexity as HFL-Net [9]. Compared to other SOTA models, UniHOPE has a moderate model size and running speeds, enabling real-time applications.

B. Implementation Details

Scene Division. Following [19], the thresholds for RRE and RTE in grasping label preparation are 5° and 10mm, respectively. An image is categorized into the hand-only scenes, if determined as non-grasping, otherwise hand-object scenes. The numbers of samples in the two scenes are shown in Tab. I. Note that although FreiHAND [23] contains a small number of images interacting with objects in both training and test sets, it cannot be divided due to the lack of object annotations.

Generative De-occluder. We adopt the officially-released pre-trained weights from [12], which fine-tunes ControlNet with synthetic hand images [1, 22]. The hand-object mask is obtained by applying dilation on the render mask of the 3D hand and object to ensure the hand-object region is covered for repainting. Then, we crop the original input image in the training set centered on the hand-object region and resize it to 512×512. The hand-object image and the hand-object mask are fed into the inpainting Stable Diffusion model, conditioned by the hand depth map. Besides, we adopt the positive prompt "a hand grasping gesture, indoor, in the lab" for image generation from the two laboratory benchmarks [2, 4], and the negative prompt is similar to the one in [12]. During inference, the number of reverse steps for DDIM is set to 50 by default.

Network Structure. (i) **Backbone**: Following [9], we adopt ResNet50 [6] as the backbone to extract features from the input image, in which a dual stream structure is adopted to relieve the competition between hand features and object features. (ii) **Hand Encoder**: The hand encoder takes $\mathbf{F}^{OH}$ as input, first using an hourglass network [13] to regress a feature map and the heatmap of 2D hand joints. Then, they are fused via a convolution layer and an element-wise addition, followed by four residual blocks to yield a 1024-dimensional vector. (iii) **MANO Decoder**: It consists of two fully connected layers to predict the hand pose and

Methods	HandOccNet [14]	MobRecon [3]	H2ONet [18]	SimpleHand [21]	Liu *et al.* [11]	Keypoint Trans. [5]	HFL-Net [9]	H2ONet + HFL-Net	HandOccNet + HFL-Net	Ours
FPS	48	78	62	41	51	33	43	36	30	44
FLOPs	15.48G	0.46G	0.74G	9.96G	39.44G	12.66G	10.01G	0.77G / 10.04G	15.51G / 10.04G	10.04G
# Param.	37.22M	8.23M	25.88M	48.89M	34.48M	52.79M	46.08M	72.26M	83.60M	46.38M

Table H. Efficiency comparison with previous methods. Note that FLOPs for the "A+B" methods depend on the predicted grasping status, therefore reported as "FLOPs of (classifier + A) / FLOPs of (classifier + B)".

Datasets (splits)	*Training Set*			*Test Set*		
	All Scenes	Hand-Only Scene	Hand-Object Scene	All Scenes	Hand-Only Scene	Hand-Object Scene
DexYCB "S0"	401,507	153,210	248,297	78,768	30,848	47,920
DexYCB "S1"	351,943	138,775	213,168	104,128	36,912	67,216
DexYCB "S3"	376,374	145,051	231,323	76,360	29,912	46,448
HO3D	66,034	5,595	60,439	11,524	2,971	8,553
FreiHAND	130,240	N/A	N/A	3,960	N/A	N/A

Table I. Number of samples in hand-only/hand-object scenes for different datasets (splits).

shape parameters of the MANO model from the feature produced by the hand encoder. (iv) **Object Decoder**: Following [9], the feature after RoIAlign from the hand branch is fused with the one from the object branch through a cross-attention layer, to enhance the object feature learning. The fused feature is then forwarded through six convolutional layers to predict the 2D projections of the 3D object corner keypoints and corresponding confidence. In testing, the object pose is computed by the Perspective-n-Point (PnP) algorithm [8] using the correspondence between the predicted 2D and the original 3D keypoints on the object mesh.

Training Details. Following [9], we perform data augmentation on the training samples, including random scaling ($\pm 20\%$), rotating ($\pm 180°$), translating ($\pm 10\%$), and color jittering ($\pm 50\%$). Our training process consists of two stages. In the first stage, the de-occluded images are incorporated into training without the feature enhancement loss for 30 epochs to first adapt the model to the domain of the generated data. In the second stage, the network is additionally supervised by the enhancement constraints between the image pairs for another 40 epochs under the same setting.

C. Limitations and Future Work

Limitations. Though we are able to predict the grasping status of unseen objects, the performance of their pose estimation tends to degrade when the object shape/appearance varies largely, due to the limited object categories in the training data. Besides, despite being provided in most existing public benchmarks, the object annotations are lacking in certain datasets, limiting the applicability of our approach as they are required for scene division and inpainting masks.

Future Work. To improve the model's generalizability towards unseen objects, a promising direction is to utilize the knowledge prior from the various vision foundation models [7, 10, 15], which demonstrated remarkable performance in zero-shot scenarios. Another approach that we are considering for improving the model's generalizability is to train on large-scale synthetic data by leveraging diffusion models [16, 19] or large language models [17].

References

[1] Synthesis AI. Static gestures dataset, data retrieved from synthesis ai. https://synthesis.ai/static-gestures-dataset/, 2023. 1, 7

[2] 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. 7

[3] 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. 1, 6, 7, 8

[4] 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, 7

[5] 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*, pages 11090–11100, 2022. 6, 7, 8

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 7

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 8

[8] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua.

EPnP: An accurate O(n) solution to the PnP problem. *IJCV*, 81:155–166, 2009. 8

[9] Zhifeng Lin, Changxing Ding, Huan Yao, Zengsheng Kuang, and Shaoli Huang. Harmonious feature learning for interactive hand-object pose estimation. In *CVPR*, pages 12989–12998, 2023. 1, 6, 7, 8

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2024. 8

[11] 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. 6, 7, 8

[12] Wenquan Lu, Yufei Xu, Jing Zhang, Chaoyue Wang, and Dacheng Tao. HandRefiner: Refining malformed hands in generated images by diffusion-based conditional inpainting. In *ACM MM*, 2024. 1, 7

[13] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer International Publishing, 2016. 7

[14] 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. 6, 7, 8

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 8

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, pages 6000–6010, 2017. 8

[17] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6D pose estimation and tracking of novel objects. In *CVPR*, pages 17868–17879, 2024. 8

[18] 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. 6, 7, 8

[19] Hao Xu, Haipeng Li, Yinqiao Wang, Shuaicheng Liu, and Chi-Wing Fu. HandBooster: Boosting 3D hand-mesh reconstruction by conditional synthesis and sampling of hand-object interactions. In *CVPR*, pages 10159–10169, 2024. 7, 8

[20] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 1

[21] Zhishan Zhou, Shihao Zhou, Zhi Lv, Minqiang Zou, Yao Tang, and Jiajun Liang. A simple baseline for efficient hand mesh reconstruction. In *CVPR*, pages 1367–1376, 2024. 6, 7, 8

[22] Christian Zimmermann and Thomas Brox. Learning to estimate 3D hand pose from single RGB images. In *ICCV*, pages 4903–4911, 2017. 1, 7

[23] 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. 7



Figure E. Qualitative comparison between UniHOPE and SOTA HPE/HOPE methods across hand-only/hand-object scenarios in DexYCB ("S3" split), in which all the grasping objects are unseen during training.



Figure F. Qualitative comparison between UniHOPE and SOTA HPE/HOPE methods across hand-only/hand-object scenarios in HO3D. The ground truths are not publicly available.

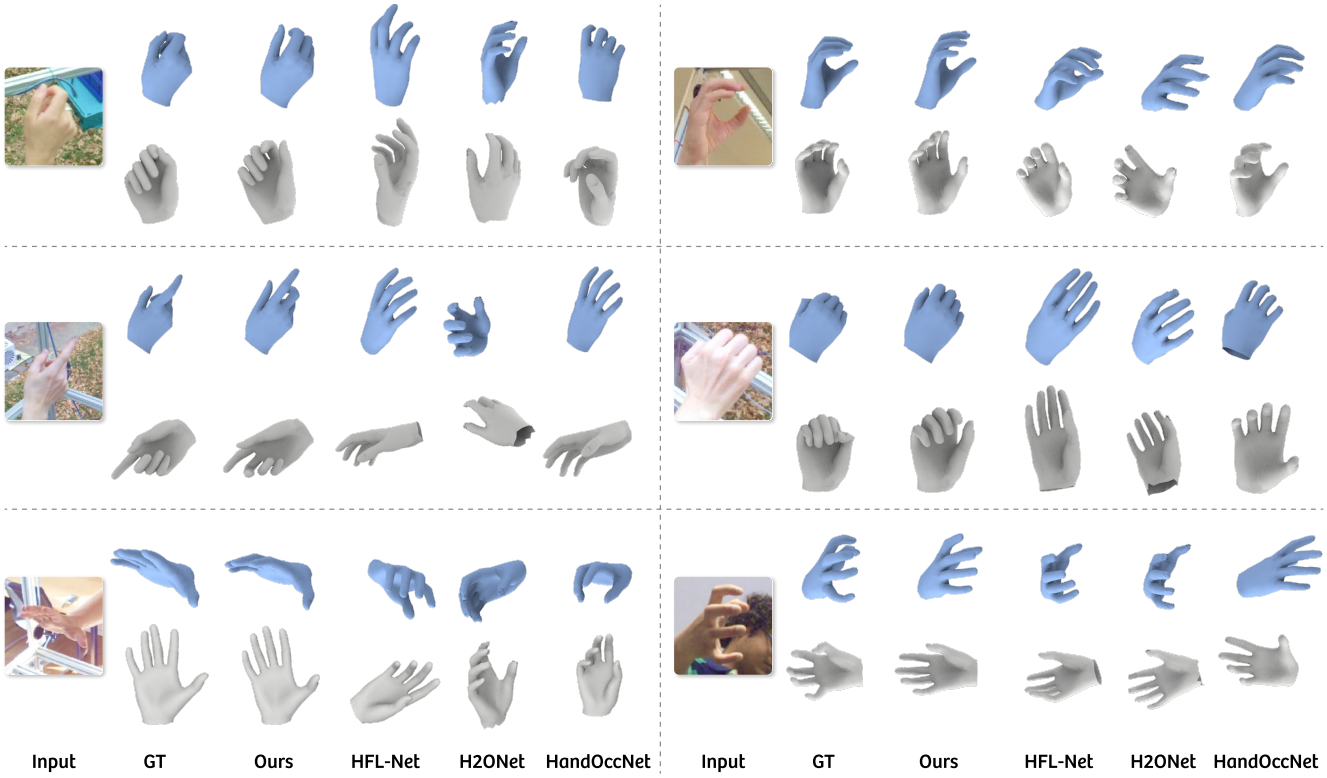


Figure G. Qualitative comparison between our method and SOTA HPE/HOPE methods on FreiHAND.

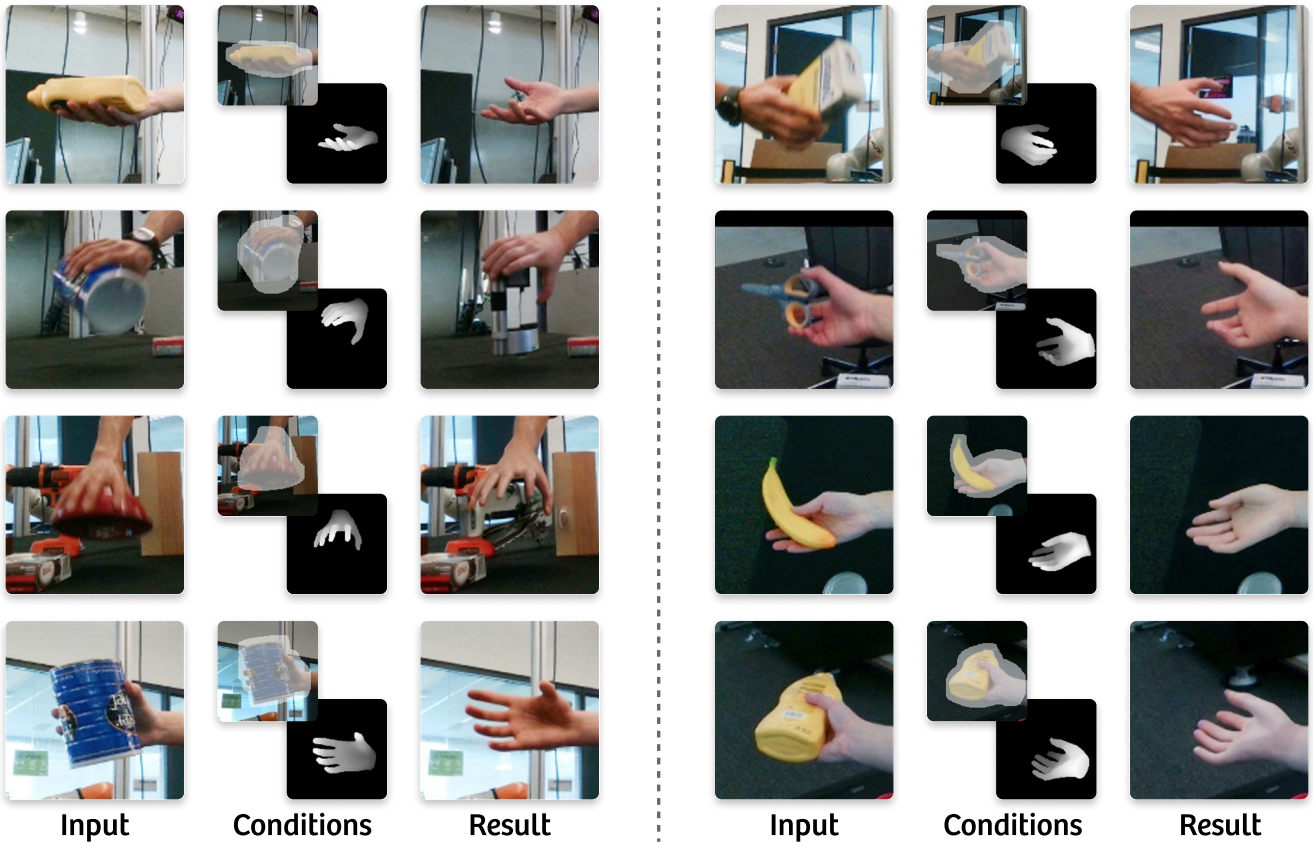


Figure H. More examples of de-occluded hand images. Note that masks are overlaid on the original image for better visualization, the actual condition for our generative de-occluder is a binary mask.

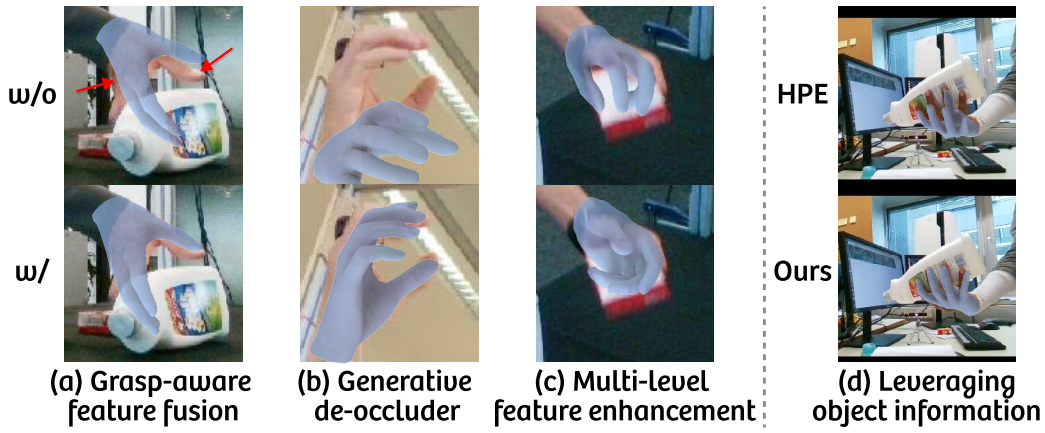


Figure I. Effects of different designs in our pipeline.

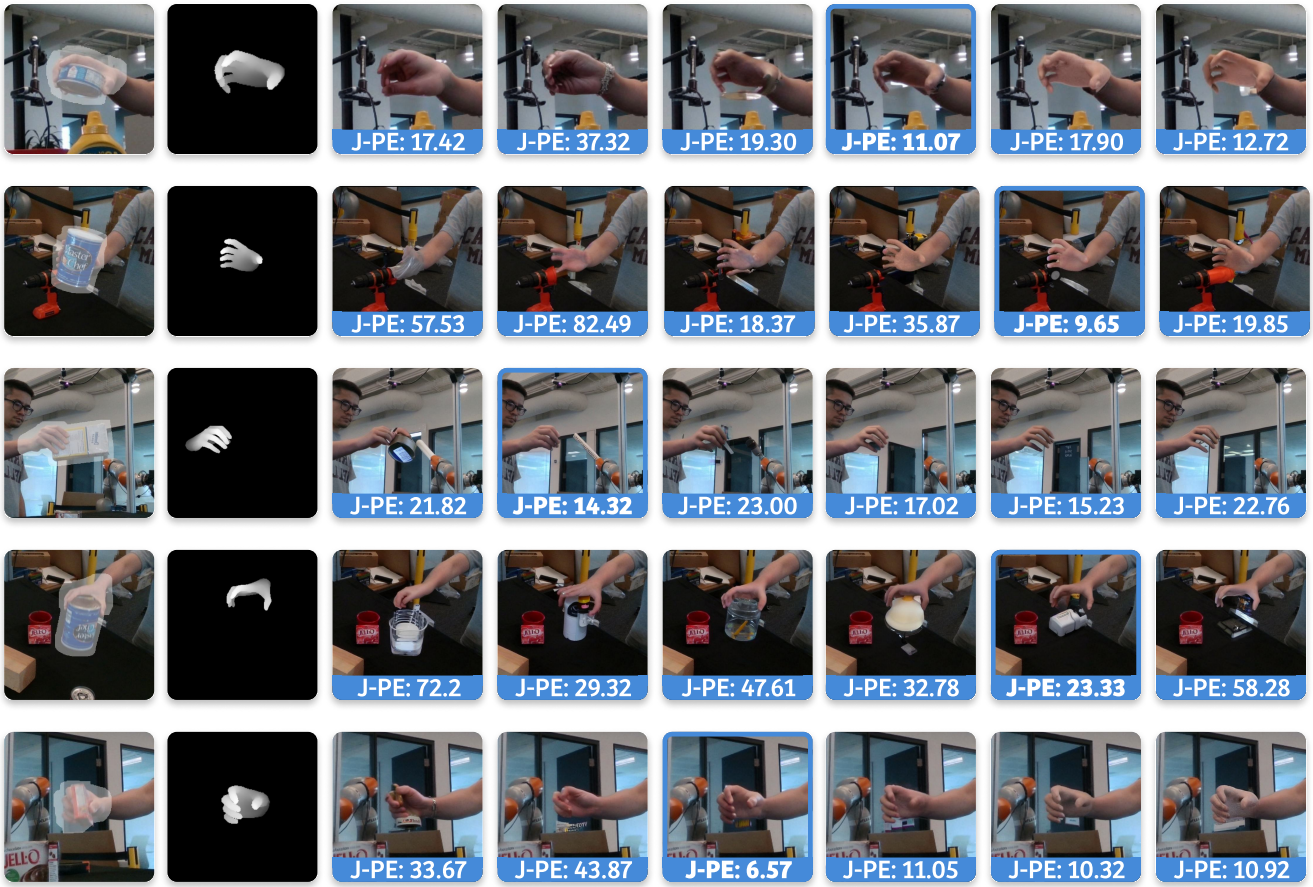


Figure J. The generated images with varying control strengths. Our adaptive strategy (metrics marked in **bold**) effectively balances fidelity and consistency.