TIGeR: Text-Instructed Generation and Refinement for Template-Free Hand-Object Interaction

Yiyao Huang[1] Zhedong Zheng[2] Ziwei Yu[1] Yaxiong Wang[3] Tze Ho Elden Tse[1] Angela Yao[1]

***Abstract*— Pre-defined 3D object templates are widely used in 3D reconstruction of hand-object interactions. However, they often require substantial manual efforts to capture or source, and inherently restrict the adaptability of models to unconstrained interaction scenarios, *e.g.*, heavily-occluded objects. To overcome this bottleneck, we propose a new Text-Instructed Generation and Refinement (TIGeR) framework, harnessing the power of intuitive text-driven priors to steer the object shape refinement and pose estimation. We use a two-stage framework: a text-instructed prior generation and vision-guided refinement. As the name implies, we first leverage off-the-shelf models to generate shape priors according to the text description without tedious 3D crafting. Considering the geometric gap between the synthesized prototype and the real object interacted with the hand, we further calibrate the synthesized prototype via 2D-3D collaborative attention. TIGeR achieves competitive performance, *i.e.*, 1.979 and 5.468 object Chamfer distance on the widely-used Dex-YCB and Obman datasets, respectively, surpassing existing template-free methods. Notably, the proposed framework shows robustness to occlusion, while maintaining compatibility with heterogeneous prior sources, *e.g.*, retrieved hand-crafted prototypes, in practical deployment scenarios.**

I. Introduction

In this paper, we study 3D reconstruction of hand-object interactions in a monocular scene. Given a single-view RGB image containing interactive behavior, we predict the 3D point clouds of hands and objects. This endeavor is crucial for enabling robots to comprehend and interact with the environment in a human-like manner, which serves as a key technology for applications, *e.g.*, Mobile ALOHA [1]. The undertaking necessitates a profound understanding of the input image and leverages the inherent 3D geometric structure priors of hands and objects to enhance the reconstruction quality. Since objects manipulated by hand have varied shapes, it is relatively challenging to obtain 3D prior knowledge of target object shapes. Some works, dubbed template-based methods [2], [3], [4], directly apply predefined object templates, *e.g.*, hand-crafted meshes, to hand-object interaction tasks. For instance, some researchers [2] resort to ground-truth object templates from the YCB dataset [5], and only need to estimate 6D-pose of the given template to match input images. However, 3D object templates are usually inaccessible in real-world scenarios. Different from template-based methods, other studies [6], [7], [8], [9], [10], referred to as template-free methods, recover UV maps or SDF representations from input RGB images. However, this line of methods typically suffers from self-occlusion by hand and thus fails to complete the entire object.

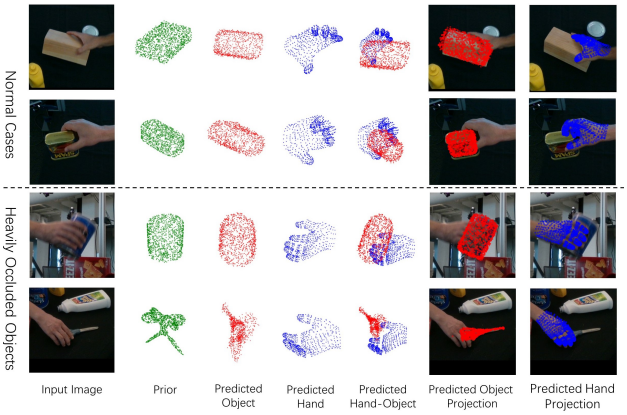


Fig. 1: Here we show the input images, generated shape priors, predicted object and hand point clouds, and the corresponding 2D projections. We could observe that the shape priors provide the common object geometry, which eases the further shape alignment. The proposed method, thus, achieves competitive reconstruction, especially for the heavy occlusions (**bottom**).

Inspired by the high-fidelity generation capabilities of cross-modal systems (particularly text-to-3D [11] and image-to-3D synthesis [12], [13], [14]), we posit that a critical research question remains underexplored: Can synthesized 3D models function as viable foundational priors to encode generalized knowledge for open-world interaction scenarios? As an early attempt to address this problem, we propose a Text-Instructed Generation and Refinement (TIGeR) framework that not only explores the prior generation pipeline but further bridges the gap between the generated prior and real-world observations. In particular, our framework consists of two sequential stages: text-instructed prior generation and vision-guided refinement. Given a hand-object interaction image, we first apply a large multimodal question-answering (QA) model to obtain the description of the target object, and then leverage the cross-modal generative models to craft the corresponding shape prior. Next, we introduce a 2D-3D collaborative attention to fuse the 3D features of the shape prior and the 2D features of the input image. Based on the fused features, our model further refines point clouds to match the target object with geometric variants, if any. Finally, TIGeR involves the hand estimation, to co-optimize hand poses, hand meshes, and translations of both hand and objects. Our method establishes

[1] Yiyao Huang, Ziwei Yu, Tze Ho Elden Tse, and Angela Yao are with National University of Singapore, Singapore 117417 {e1322639,yuziwei}@u.nus.edu, eldentseb@gmail.com, ayao@comp.nus.edu.sg

[2] Zhedong Zheng is with University of Macau, Macau, China 999078 zhedongzheng@um.edu.mo

[3] Yaxiong Wang is with Hefei University of Technology, Hefei, China 230009 wangyx@hfut.edu.cn

correspondences between 3D point clouds and 2D images, enabling alignment for real hand-object interaction data. The entire process does not require any 3D template annotations, easing pre-requisites for real-world scenarios. Therefore, our contributions are as follows:

- **Template-free Framework.** Different from existing works demanding a pre-defined object template, we introduce a Text-Instructed Generation and Refinement (TIGeR) framework to improve the scalability and ease the prerequisites for 3D hand-object interaction reconstruction. Inspired by the recent success of text-based 3D object generation, we borrow the strength of text-driven prior to replace the hand-crafted template, and validate the feasibility.
- **Cookbook for Prior Refinement.** Given the gap between the 3D prior and the real object in the photo, we introduce an attention-based paradigm to further register the object according to the visual cues. In particular, we integrate both 2D and 3D features via 2D-3D collaborative attention module, simultaneously performing shape refinement and object registration.
- **Competitive and Robust Performance.** We evaluate our framework on two large-scale hand-object interaction datasets, *i.e.*, Dex-YCB[15] and Obman [8], surpassing competitive template-free approaches. Moreover, our method is robust against the common hand-occluded cases and also scalable to other prototype sources, *e.g.*, retrieved hand-crafted samples.

II. Related Work

3D hand pose and shape estimation. Hand pose estimation methodologies have evolved through three technical paradigms. Early learning-based approaches [16], [17], [18], [19] employ direct 3D keypoint regression from RGB inputs, and produce anatomically inconsistent surfaces that hinder downstream applications. This limitation motivates parametric modeling [20], [21], *e.g.*, MANO [21] establishing a kinematic hand model. Besides, non-parametric paradigms [22], [23] circumvent shape space constraints through vertex-level prediction, employing disentangled autoencoders to isolate pose dynamics from background interference. Recent approaches integrate neural texture representations via UV mapping [24], [9] and geometric attention mechanisms in transformer architectures [25], [26], [27], [28], enabling joint optimization of skeletal pose and surface deformation.

3D object reconstruction. Early voxel-based approaches [29], [30], [31] establish grid-form representations, yet remain constrained by cubic memory complexity. Subsequent approaches transitioned to point cloud representations [32], [33], [34], [35], employing graph-based aggregation modules to model local geometric structures. The field advances through implicit surface representations, with Park *et al.* [36] pioneering memory-efficient shape encoding via continuous signed distance fields. Concurrent surface deformation strategies emerge, including FoldingNet's parameterized grid transformation [37] and AtlasNet's MLP-driven mesh generation from primitive patches [38]. Beyond geometric reconstruction, some works on pose estimation [39], [40], [41], [42] fuse RGB-D data to recover 6D object poses. Modern frameworks [43], [44] instead perform cross-modal feature alignment, establishing geometric correspondences between 2D projections and 3D assets to derive pose parameters.

3D hand-object interaction. Recent works primarily fall into two categories: template-based and template-free approaches. Template-based methods [2], [45], [46] rely on RGB images paired with 3D object templates, leveraging multi-modal inputs for enhanced precision. Traditional pipelines [46] employ hand pose regression followed by SfM initialization and refinement. Recent implementations extract global image features to estimate MANO parameters and 6D object poses [2], while hybrid architectures combining single- and dual-stream backbones through ROIAlign operations [45]. Template-free approaches [8], [6], [47] reconstruct geometry directly from RGB images without explicit shape priors. Early methods deform a parametric sphere using global features [8], [38], while recent techniques employ SDF decoders to integrate visual and pose cues [6], [47]. Despite their applicability to real-world data, these methods remain sensitive to image degradation and occlusion-induced ambiguities. Our framework mitigates these issues by introducing text-guided shape priors from multimodal generative models, enabling detailed single-view reconstruction via semantic-geometric alignment.

III. Method

Given a hand-object interaction image, our task is to reconstruct both 3D hand and object without relying on hand-crafted templates, which is usually inaccessible in real-world scenarios. Our framework contains two primary stages, *i.e.*, text-instructed prior generation (see Fig. 2), and vision-guided refinement (see Fig. 3). During the first prior generation stage, we leverage off-the-shelf generative models to craft coarse shape prior $\bar{V}$ from the input image I. While this prior captures general semantic structure, it often lacks fine-grained geometric details aligned with the input image. In the second vision-guided refinement stage, we intend to explicitly reduces the geometric discrepancy between the generated prior and the actual object in the image. In particular, we extract 2D visual features from the input image I and 3D geometric features from the prior $\bar{V}$ and integrate both 2D and 3D features by leveraging 2D-3D collaborative attention modules. The fused features are then decoded into a refined point cloud $\tilde{V}$. Finally, we perform joint optimization of both the 3D hand and object to estimate their poses and output the final hand-object point clouds. In the following subsections, we elaborate two stages respectively.

A. Text-instructed Prior Generation

Prior Generation. As shown in Figure 2, our generation pipeline contains three phases: (1) Captioning, (2) Text-to-Image Generation, and (3) Image-to-Point Cloud Generation. We intend to obtain category information of target objects in the first phase. To this end, we query a pre-trained image caption model, *e.g.*, InstrucBLIP [48], using the prompt *"What is being held by hand?"* The output text follows a specific format: *"In this image, the hand is holding a [The category*

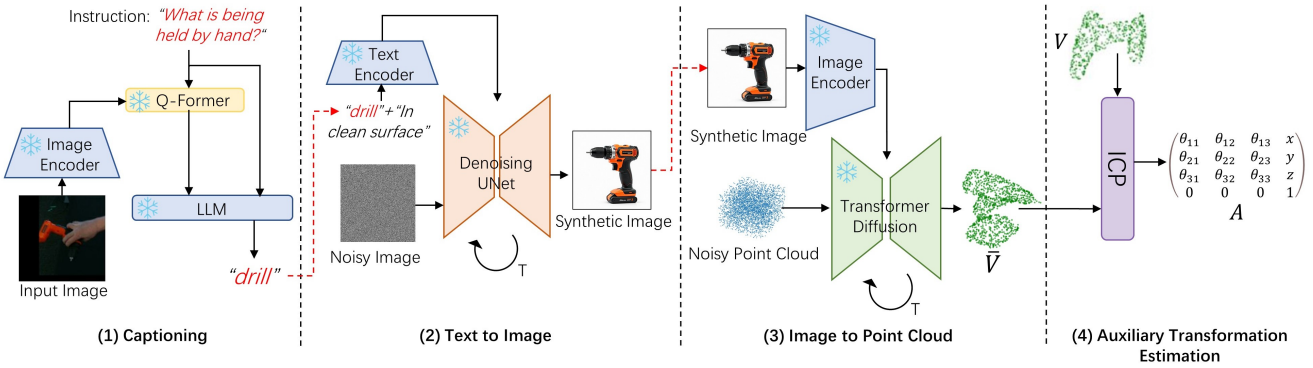


Fig. 2: A brief text-instructed prior generation pipeline. **(1) Captioning.** Given an input image depicting hand-object interaction, we first identify the occluded object by querying a multimodal large-scale model with the prompt: "What is being held by hand?" **(2) Text-to-Image Generation** Using the generated caption, we condition a diffusion model to synthesize a canonical view of the object without occlusions. **(3) Image-to-Point Cloud Generation.** Finally, we employ an off-the-shelf 2D-to-3D lifting model to generate a 3D shape prior from the synthetic image. **(4) Auxiliary Transformation Estimation**. We estimate the auxiliary transformation A between the shape prior and the ground-truth point cloud by Iterative Closet Point (ICP).

name of the object]." Then, in the Text-to-Image Generation phase, we feed the structured text prompt *"A [The category name of the object] in a clean surface"* into a text-to-image generator, *e.g.*, Diffusion Model [49], to obtain a synthetic image with a clear background that only contains the target object. Next, we leverage the off-the-shelf image-to-point cloud model, *e.g.*, Point-E [12], to generate the coarse-grained point cloud $\bar{V}$ as the 3D shape prior. Lastly, we apply Iterative Closest Point (ICP) to find the optimal transformation for $\bar{V}$. **In this work, we do not pursue an optimal 3D prior but focus on validating the feasibility of the text-driven prior to replace the hand-crafted template.**

Auxiliary Transformation Estimation. To facilitate the model training, we also estimate the optimal transformation between the generated prior and the ground-truth mesh in the training set. In this way, we could have a pseudo one-to-one correlation during training to stabilize the model training in the early stage. Given the synthesized $\bar{V}$ and the ground-truth mesh V, we derive the optimal transformation matrix A by Iterative Closest Point (ICP):

$$A = \underset{A}{\operatorname{argmin}} \|V - A\bar{V}\|_2^2. \tag{1}$$

Given the predicted transformation A, we could have a pseudo one-to-one mapping between synthesized $\bar{V}$ and the ground-truth mesh V as $\mathcal{J}(i) = \operatorname{argmin}_j \|V_i - A\bar{V}_j\|_2^2$. $\mathcal{J}(i)$ denotes the index of $\bar{V}_j$, which is the nearest neighbor of V_i. **We note that we do not use such estimation during inference. The auxiliary pseudo transformation is only estimated for training.**

B. Vision-guided Refinement

Shape refinement. As shown in Figure 3, we show the brief structure of our vision-guided refinement stage. Given a coarse shape prior $\bar{V}$ and an input image I, we first extract complementary 2D and 3D features through dedicated visual and geometric encoders. Since shape prior $\bar{V}$ contains category-level geometric knowledge, such as the cuboid structure of boxes, the object shape geometric encoder processes $\bar{V}$ through two hierarchical layers, producing local features F_g^1 and F_g^2. Similarly, we obtain multi-resolution visual features from the input image I via the object shape visual encoder, yielding local visual features F_v^1, F_v^2 and global feature F_{vg} via average pooling. To align the shape prior $\bar{V}$ with the hand-object interaction scene, we propose a cross-modal feature fusion approach that establishes correspondences between 3D patches and 2D image regions. The fusion process begins by repeating and concatenating the global visual feature F_{vg} with each 3D patch's geometric features to form an initial fused representation. The fused representation is then processed by MLPs followed by softmax to generate attention weights $W_l, l \in \{1, 2\}$, which identify the relevant image regions for each 3D patch. W_l are applied to the visual features F_{wv}^l. We finally concatenate F_{wv}^l with F_g^l to get the fused feature F_{fused}^l, which contains both precise information from 3D patches and rich visual cues from the corresponding image regions. Given the fused local feature F_{fused}^l, the object shape geometric decoder predicts adjusted 3D coordinates to align the shape prior with the interaction scene. The decoding, following U-net [50] style, consists of two processes. First, F_{fused}^2 is processed through MLPs and interpolated to generate features for intermediate points. The F_{fused}^1 is then concatenated with these intermediate features and passed through additional MLPs, followed by linear interpolation to complete features for all remaining points. Finally, the network regresses 3D coordinates for every vertex to produce the aligned object point cloud $\tilde{V}$.

Pose estimation. Simultaneously, we predict the object center location and the hand pose through two independent object and hand pose visual encoders, respectively. As shown in the bottom of Figure 3, we apply the hand pose visual encoder to extract 21 feature maps from the input image I. Then, we obtain the uvd (u+v+depth) location of the max activation values in every heatmap as 21 hand key points. Given the camera intrinsic, the uvd coordinates can be transformed into 3D positions $\tilde{C}_h$ in the world coordinate system. We apply Inverse kinematics [51] to convert $\tilde{C}_h$ into MANO parameters, which are then provided to the MANO model to get the hand vertices $\tilde{H}$. Similarly, given the input image,

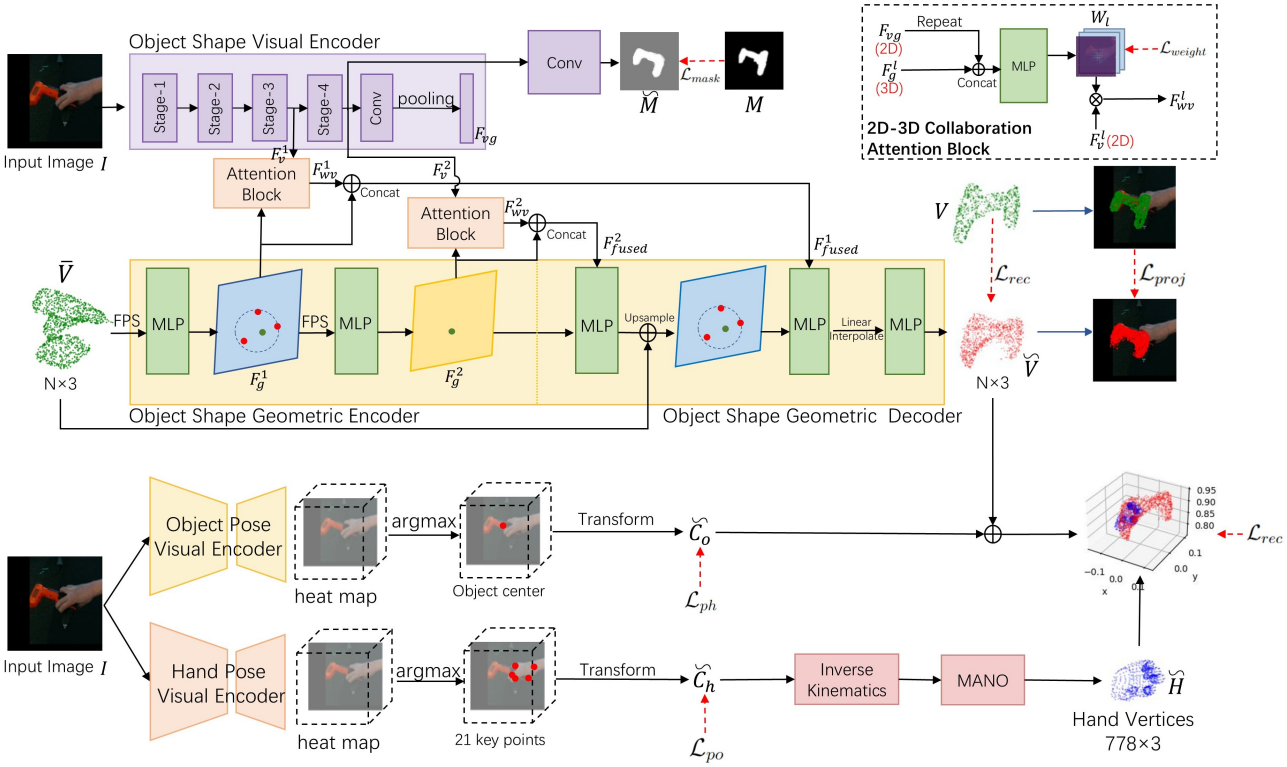


Fig. 3: Overview of vision-guided refinement stage. **Top:** Given the text-driven shape prior $\bar{V}$ and the RGB image I, we extract the 2D visual feature via object shape visual encoder and the 3D geometric feature via object shape geometric encoder. Then we apply 2D-3D collaboration attention blocks (*top right*) to fuse the visual feature and the geometric feature. The fused features are then fed to the object shape geometric decoder to predict the object shape $\tilde{V}$. **Bottom:** Given the input image I, we estimate the object center in the image and hand poses, *i.e.*, 21 key points. On one hand, the input image is fed to the object pose visual encoder, which does not share weight with the object shape visual encoder, to obtain the center estimation. On the other hand, we apply the hand pose visual encoder to predict 21 key points. We then manipulate the MANO model to reconstruct the hand $\tilde{H}$. Finally, we fuse the object point cloud and hand point cloud supervised by $\mathcal{L}_{rec}$.

we apply the object pose visual encoder to extract the object heat map, and then obtain the index for the point with the maximum activation value. Then we transform the point index into the 3D coordinates of the object center $\tilde{C}_o$. We translate the refined object $\tilde{V}$ to the predicted center, and compose the reconstructed object and hand as the final output.

Optimization objectives. To facilitate reconstructing the geometric shape of the target object in the early training, we introduce several auxiliary tasks. For instance, we leverage the pseudo one-to-one mapping $\mathcal{J}(i)$ (defined in Section III-A) from the point index of the target object V to the point index of the shape prior $\bar{V}$ to supervise the intermediate attention weight W_2, which is in the second 2D-3D Collaboration Attention Block as:

$$\mathcal{L}_{weight} = \frac{1}{|V|} \sum_{i,j=\mathcal{J}(i)} \|\phi(V_i) - argmax(W_2(j))\|_2^2, \quad (2)$$

where ϕ denotes the 2D projection of the vertex in the ground-truth V. The second term is the coordinates of max activation in the corresponding heatmap W_2. Similarly, we apply the pseudo one-to-one mapping to supervise the projection of the final reconstructed object as:

$$\mathcal{L}_{proj} = \frac{1}{|V|} \sum_{i,j=\mathcal{J}(i)} \|\phi(V_i) - \phi(\tilde{V}_j)\|, \quad (3)$$

For 3D supervision, we apply the conventional group-to-group reconstruction loss via Chamfer Distance as :

$$\mathcal{L}_{rec} = \frac{1}{|\tilde{V}|} \sum_{i=1}^{|\tilde{V}|} \min_j \|\tilde{V}_i - V_j\|_2^2 + \frac{1}{|V|} \sum_{j=1}^{|V|} \min_i \|V_j - \tilde{V}_i\|_2^2. \quad (4)$$

Furthermore, we introduce foreground mask supervision as an auxiliary task to make our object shape visual encoder concentrate on the target object and mitigate the negative impact of occlusion. Specifically, we take F_v^2 as input followed by a 2D-convolutional layer, a max pooling layer and sigmoid function to estimate $\tilde{M}$ as the foreground probability. The foreground mask loss is a binary classification task as:

$$\mathcal{L}_{mask} = -\sum_i (M_i log(\tilde{M}_i) + (1 - M_i) log(1 - \tilde{M}_i)), \quad (5)$$

where M is the resized ground-truth amodal mask. For the two pose visual encoders, we introduce $\mathcal{L}_{ph}$ and $\mathcal{L}_{po}$ as L2 distance between the prediction $\tilde{C}$ and the corresponding ground truth C, which can be formulated as:

$$\mathcal{L}_{ph} = ||C_h - \tilde{C}_h||_2^2, \mathcal{L}_{po} = ||C_o - \tilde{C}_o||_2^2. \quad (6)$$

Therefore, the final loss function for the shape refinement

stage is:

$$\mathcal{L}_{registration} = \mathcal{L}_{rec}+\mathcal{L}_{mask}+\mathcal{L}_{ph}+\mathcal{L}_{po}+\lambda_{weight}\mathcal{L}_{weight}+\lambda_{pro} \quad (7)$$

Considering that $\mathcal{L}_{weight}$ and $\mathcal{L}_{proj}$ are based on the pseudo alignment, we empirically set a relatively small weight, *i.e.*, $\lambda_{weight} = 0.1$, $\lambda_{proj} = 0.01$.

IV. Experiment

Datasets. (1) DexYCB contains 582K RGB-D frames of single-hand object grasping from 8 views, providing 3D hand poses and 6D object poses for 20 YCB-Video objects [41]. Each frame includes a target object and 1–3 distractors on a black table. Following [47], we sample every 6th frame, resulting in 29K training and 5K test samples. Original 640×480 RGB images are cropped to 256×256, centered on the target. **(2) Obman** [8] is a synthetic dataset with 150K images, using 2,772 ShapeNet object meshes [52]. Hand poses and meshes are generated via GraspIt [53] and MANO [21], then rendered with LSUN [54] and ImageNet [55] backgrounds at 256×256 resolution. Following [47], we split it into 87K training and 6K test samples.

Metrics. We evaluate the quality of both hand and object reconstruction by computing the Chamfer Distance(mm) between the predicted point clouds and the ground truth point clouds. We also report F-score at 5 mm (FS_o@5) and 10 mm (FS_o@10) as thresholds for predicted object point clouds and F-score at 1 mm (FS_h@1) and 5 mm (FS_h@5) as thresholds for predicted hand point clouds.

Compared methods. We mainly compare our method with 3 competitive template-free methods. (1) Hasson *et al.* [8]: A classical template-free method that reconstructs hand and object meshes from an RGB image using global visual features decoded via AtlasNet [38] (object) and MANO layers [21] (hand). (2) AlignSDF [6]: An SDF-based approach leveraging global visual features and pose information to reconstruct hand and object surfaces using an SDF decoder. (3) gSDF [47]: Another SDF-based method utilizing local visual features from image feature maps. It predicts 3D keypoints from heatmaps and employs Inverse Kinematics [51] to refine hand pose estimation.

Implementation details. The hand-object reconstruction pipeline comprises two stages. (1) **Text-instructed prior generation**: Three pre-trained generative models are utilized to produce shape priors from RGB images: InstructBLIP [48] for captioning, FLUX-1 [49] for image synthesis, and Point-E [12] for point cloud generation. **Note that this work does not seek optimal priors but validates their effectiveness; the method supports alternative prior sources (see Section IV-B).** (2) **Vision-guided refinement**: HRNet [56] and PointNet++ [33] serve as visual and geometric backbones for shape feature extraction. ResNet-50 [57] is used for both hand and object pose encoding. The model is trained for 1,600 epochs using Adam [58] with an initial learning rate of $5e^{-5}$ on 4 NVIDIA RTX A5000 GPUs. To reduce assembly error, the full pipeline is fine-tuned for 100 epochs with $\mathcal{L}_{rec}$ at a learning rate of $1e^{-5}$, while freezing the object shape and hand pose encoders.

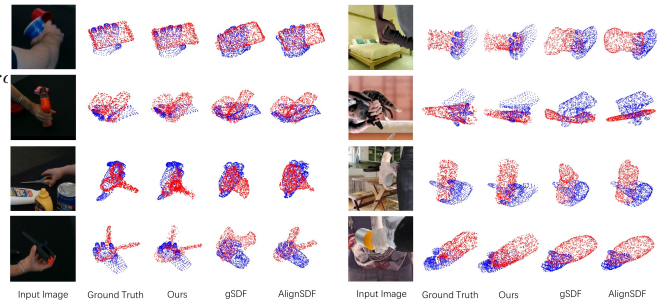


Fig. 4: Qualitative comparision of TIGeR (Ours) and prevailing template-free methods [47], [6], [8] on DexYCB (*left*) and Obman (*right*).

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

As shown in Table I, we could observe that the quality of objects reconstructed by our method surpasses the quality of objects produced by template-free SOTA methods on both the DexYCB dataset and the Obman dataset. For instance, our method has arrived at 1.979 median Chamfer Distance (CD_o) 0.292 FS_o@5 and 0.637 FS_o@10 on the DexYCB dataset, which surpasses gSDF [47] by a clear margin. We observe a similar phenomenon on the Obman dataset. Our method has achieved 5.468 CD_o, and competitive 0.199 FS_o@5 and 0.462 FS_o@10 scores. As for reconstructed hands, on both datasets, our method yields high-quality hands with the lowest median Chamfer Distance (CD_h), while yielding the highest FS_h@1 and FS_h@5. Furthermore, we show the qualitative comparison of our methods and SOTA methods in Figure 4. Our method achieves superior geometric fidelity for both simple primitives (*e.g.*, cans, boxes) by preserving sharp edges and planar surfaces, and complex articulated objects (*e.g.*, scissors, drills) through high-fidelity detail retention, whereas competitive methods exhibit significant shape distortions and topological oversimplification. Note that SDF-based methods generate uniformly distributed points in the reconstructed surface geometry, resulting in geometrically ambiguous reconstructions of articulated hand regions. This inherent uniformity inadequately captures the non-linear deformation patterns required for dexterous finger manipulation in real-world scenarios. In contrast, the proposed method leverages the straightforward kinematic-aware hand parametric model and generated object priors to ease the optimization difficulty, while preserving more interaction details.

B. Ablation Studies and Further Discussion

Comparison of the object reconstruction only. To isolate object reconstruction quality, we center-normalize both predicted and ground-truth objects by aligning their centroids to the origin. Our re-implementation of three competitive methods reveals that Hasson [8] achieves optimal CD_o (1.17/2.90) and FS_o scores when evaluated purely on object reconstruction. As shown in Table II, our method further reduces the Chamfer distance as 0.62 on DexYCB and 2.78 on Obman, surpassing Hasson by a clear margin.

Robustness against occlusions. The target objects interacted by the human are usually occluded by hands or other objects.

Dataset	Method	$CD_o \downarrow$	$FS_o@5 \uparrow$	$FS_o@10 \uparrow$	$CD_h \downarrow$	$FS_h@1 \uparrow$	$FS_h@5 \uparrow$
DexYCB	Hasson [8]	5.831	0.155	0.405	6.375	0.003	0.162
	AlignSDF [6]	2.669	0.254	0.588	2.768	0.003	0.222
	gSDF [47]	2.769	0.258	0.591	2.770	0.003	0.222
	TIGeR (Ours)	**1.979**	**0.292**	**0.637**	**1.132**	**0.008**	**0.413**
Obman	Hasson19 [8]*	-	-	-	-	-	-
	AlignSDF [6]	5.584	0.203	0.476	2.117	0.004	0.248
	gSDF [47]	5.626	**0.207**	**0.482**	2.116	0.004	0.249
	TIGeR (Ours)	**5.468**	0.199	0.462	**0.787**	**0.013**	**0.537**

TABLE I: Quantitative results of hand-object reconstruction performance on DexYCB and Obman. *: We re-implement the official code but the method does not converge when involving both hand and object.

Method	DexYCB			Obman		
	$CD_o \downarrow$	$FS_o@5 \uparrow$	$FS_o@10 \uparrow$	$CD_o \downarrow$	$FS_o@5 \uparrow$	$FS_o@10 \uparrow$
Hasson [8]	1.17	0.36	0.78	2.90	0.27	0.61
AlignSDF [6]	1.41	0.37	0.73	3.65	0.24	0.54
gSDF [47]	1.53	0.36	0.72	3.88	0.23	0.53
TIGeR (Ours)	**0.62**	**0.54**	**0.90**	**2.78**	**0.30**	**0.63**

TABLE II: Comparison of reconstructed object **only**. We have centralized all objects for a fair comparison.

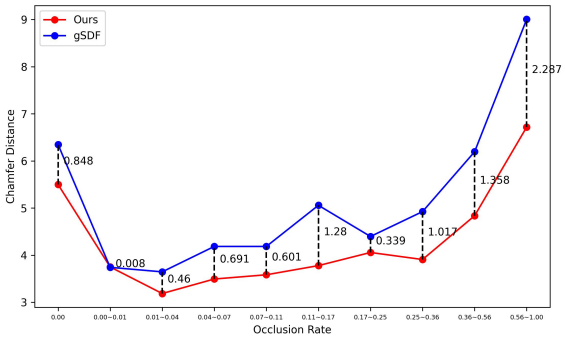


Fig. 5: Comparison between ours and the competitive gSDF against the occlusion. We could observe that our method has achieved lower Chamfer distance than gSDF in all ranges of occlusion rate, especially for heavy occlusion.

We analyze the relationship between the reconstruction quality of objects and the degree of occlusion. We employ the ground truth amodal mask M_{amodal} and visible mask $M_{visible}$ of the target objects to measure the occlusion rate of testing samples as $R_{occlusion} = 1 - \frac{Area(M_{visible})+1}{Area(M_{amodal})+1}$, where Area$(\cdot)$ denotes the area of the foreground in the corresponding mask. We split the samples into 10 equal-numbered groups according to their occlusion rate. In Figure 5, we report the median Chamfer Distance between the predicted point clouds and the ground truth point clouds for every group. As the increasing occlusion rate, the proposed method yields a clear margin towards competitive gSDF. As shown in Figure 6, we visualize some samples with severe occlusion compared to gSDF. This robustness stems from our shape prior to provide geometric cues: (1) For simple geometric objects, *e.g.*, cans and boxes, the prior effectively preserves sharp edges and planar surfaces even under heavy occlusion; (2) For complex articulated objects, *e.g.*, scissors, the prior maintains proper handle and blade geometry. We highlight the discrepancy in Figure 6 with green circles.

Study of prior quality. We show intermediate results of the first text-instructed prior generation stage in Figure 7. We observe that our generation result gradually approaches the target object. We also quantitatively study the generated prior

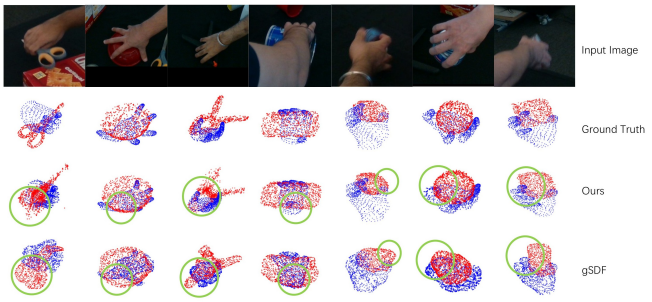


Fig. 6: Qualitative comparison between our method and gSDF under severe occlusion scenarios. The green circles highlight the prediction discrepancy.

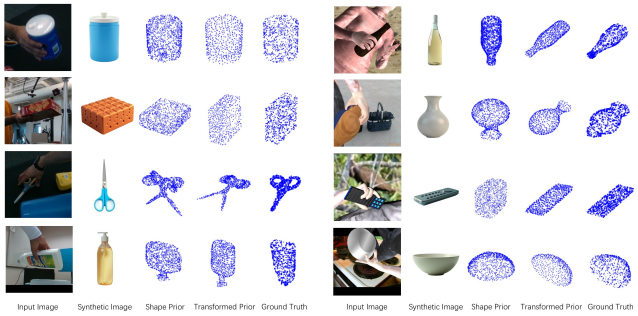
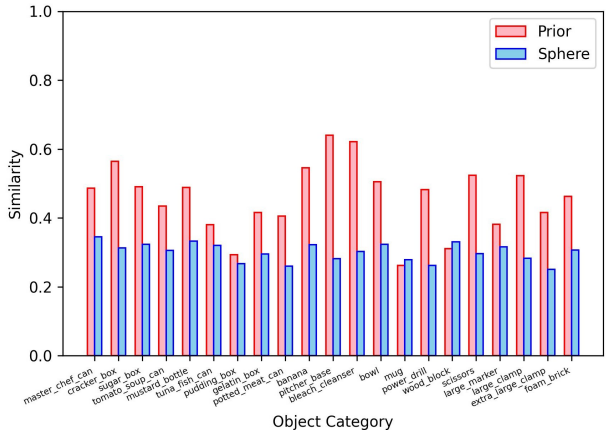


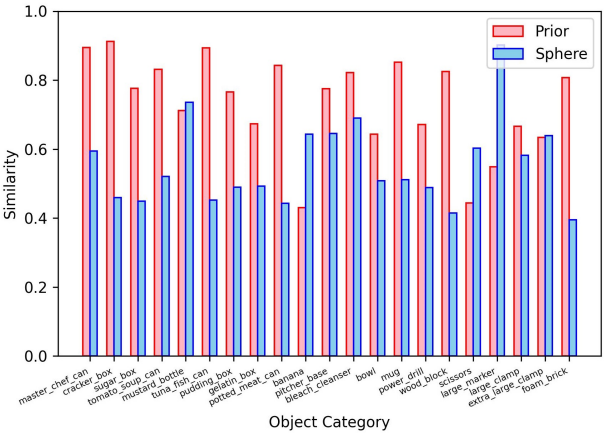
Fig. 7: Here we show the intermediate during generation, including the text-to-image result (*i.e.*, synthetic image), image-to-3D result (*i.e.*, shape piror), and the pseudo transformation (*i.e.*, transformed priors) and ground-truth object on the training set of DexYCB (*left*) and Obman (*right*).

quality by comparing it with the commonly-used unit sphere. In particular, we adopt DGCNN [59] to extract perceptual features, and calculate the feature similarity with the ground-truth. As shown in Figure 8a, we find that the generated prior easily surpasses the sphere unit. We further apply the pseudo transformation to both our prior and the sphere in Figure 8b. The proposed method yields a higher similarity score among all subcategories.

Scalability to different prior sources. To validate that our framework is compatible with different prior sources, we adopt the retrieved images to replace the synthetic images during the prior generation. As shown in Table IIIa, we observe that priors from synthetic images also perform well, surpassing the baseline with unit sphere by a clear margin. We further analyze the performance of different object categories



(a) Without Transformation



(b) With Pseudo Transformation

Fig. 8: Similarity between our prior and ground-truth template (red), sphere and ground-truth template (blue) in terms of different object categories on the DexYCB training set. **Higher is better.** We observe that whether the transform is performed or not, the generated prior is more similar to the ground truth than the widely-used sphere initialization, facilitating the optimization.

(*i.e.*, boxes, cans, bottles, others on DexYCB) in Table IIIc. Except for the sphere-alike 'can' objects, our prior usually achieves lower median Chamfer Distance than the unit sphere.

Effect of two vision-guided losses. We introduce two vision-guided loss terms, *i.e.*, $\mathcal{L}_{weight}$ and $\mathcal{L}_{proj}$, to regulate attention mechanisms in correlating 3D prior patches with 2D image regions. Our ablation studies on the DexYCB dataset (Table IIIb) validate the effectiveness of $\mathcal{L}_{weight}$ and $\mathcal{L}_{proj}$. When training without $\mathcal{L}_{weight}$, we observe a significant increase in the median Chamfer distance between the centered predicted point clouds and ground truth. Similarly, removing $\mathcal{L}_{proj}$ leads to a 0.7 increase in this metric. We further visualize attention maps in Figure 9 for 512 points with and without these two losses. Without $\mathcal{L}_{weight}$, all query points incorrectly focus on the zero (u,v) region, forcing the decoder to use uninformative top-left corner features for coordinate prediction. Without $\mathcal{L}_{proj}$, attention becomes overly concentrated at the object center, neglecting edge features. Joint application of both losses enables spatially distributed attention, providing the decoder with comprehensive local visual cues for decoding.

TABLE III: **Ablation studies. (a)** We adopt different prior sources for comparison. **(b)** We study the effect of two projection losses. **(c)** We study the performance of four different sub-categories based on the proposed prior and commonly-used sphere prototype.

(a)

Source of prior	$CD_o \downarrow$	$FS_o@5 \uparrow$	$FS_o@10 \uparrow$
Unit Spheres	0.67	0.53	0.88
Retrieval Images	0.65	0.54	0.90
Synthetic Images	**0.62**	**0.55**	**0.90**

(b)

$\mathcal{L}_{weight}$	$\mathcal{L}_{proj}$	$CD_o \downarrow$	$FS_o@5 \uparrow$	$FS_o@10 \uparrow$
✗	✓	3.88	0.24	0.57
✓	✗	1.32	0.38	0.77
✓	✓	**0.62**	**0.55**	**0.90**

(c)

Category	Prototype Prior	Prototype Sphere	$Sim. \uparrow$	$CD_o \downarrow$
Boxes	✓		**0.775**	**0.585**
		✓	0.496	0.635
Cans	✓		**0.707**	0.585
		✓	0.563	**0.572**
Bottles	✓		**0.743**	**0.639**
		✓	0.639	0.705
Others	✓		**0.712**	**0.678**
		✓	0.561	0.780

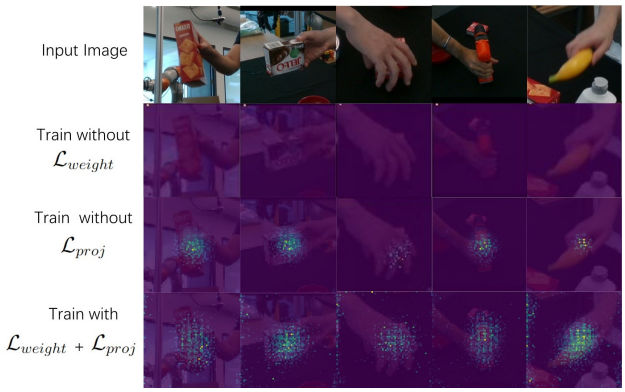


Fig. 9: Attention maps of 512 query points on the shape prior. The bright areas indicate the high probability to be an object.

Limitation. TIGeR inherits the constraints from off-the-shelf generative models. Specifically, our current implementation struggles with objects exhibiting significant intra-class shape diversity under functional states, *e.g.*, modeling both open and closed configurations of scissors. This limitation stems from existing generative priors prioritizing inter-class discriminability over fine-grained state variations, occasionally leading to ambiguous geometric reconstructions in dynamic interaction scenarios. Based on more future work on fine-grained 3D generation, our method would further improve the scalability.

V. Conclusion

In this paper, we present Text-Instructed Generation and Refinement (TIGeR) for 3D hand-object interaction estimation that addresses the scalability of template-based approaches. By synergizing cross-modal generative model with geometric refinement, TIGeR eliminates reliance on hand-crafted templates while maintaining interpretability through its two-stage design, *i.e.*, text-instructed prior generation and vision-guided refinement. We also provide a cookbook to complete prior registration and shape deformation using attention blocks to fuse the local 2D visual features and 3D geometric features. Extensive evaluations on two widely-used benchmarks, *i.e.*, DexYCB and Obman, verify the effectiveness of the generated 3D prior, outperforming existing methods by 0.034 in F-score@5 while reducing shape reconstruction Chamfer Distance by 0.69. The framework's generalizability is further evidenced by its robustness against occlusions and seamless integration with different priors, *e.g.*, retrieved priors.

REFERENCES

[1] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," *arXiv:2401.02117*, 2024.

[2] Y. Hasson, B. Tekin, F. Bogo, I. Laptev, M. Pollefeys, and C. Schmid, "Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction," in *CVPR*, 2020.

[3] Y. Hasson, G. Varol, C. Schmid, and I. Laptev, "Towards unconstrained joint hand-object reconstruction from rgb videos," in *3DV*, 2021.

[4] Z. Zheng, X. Wang, N. Zheng, and Y. Yang, "Parameter-efficient person re-identification in the 3d space," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 6, pp. 7534–7547, 2022.

[5] Z. Cao, I. Radosavovic, A. Kanazawa, and J. Malik, "Reconstructing hand-object interactions in the wild," *ICCV*, 2021.

[6] Z. Chen, Y. Hasson, C. Schmid, and I. Laptev, "AlignSDF: Pose-aligned signed distance fields for hand-object reconstruction," in *ECCV*, 2022.

[7] Y. Chen, Z. Tu, D. Kang, L. Bao, Y. Zhang, X. Zhe, R. Chen, and J. Yuan, "Model-based 3d hand reconstruction via self-supervised learning," in *CVPR*, 2021.

[8] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid, "Learning joint reconstruction of hands and manipulated objects," in *CVPR*, 2019.

[9] Z. Yu, L. Yang, Y. Xie, P. Chen, and A. Yao, "Uv-based 3d hand-object reconstruction with grasp optimization," *arXiv:2211.13429*, 2022.

[10] T. H. E. Tse, K. I. Kim, A. Leonardis, and H. J. Chang, "Collaborative learning for hand and object reconstruction with attention-guided graph convolution," in *CVPR*, 2022.

[11] H. Yi, Z. Zheng, X. Xu, and T.-s. Chua, "Progressive text-to-3d generation for automatic 3d prototyping," *ACM TOMM*, 2026.

[12] A. Nichol, H. Jun, P. Dhariwal, P. Mishkin, and M. Chen, "Point-e: A system for generating 3d point clouds from complex prompts," *arXiv:2212.08751*, 2022.

[13] J. Wang, Z. Zheng, W. Xu, and P. Liu, "Rigi: Rectifying image-to-3d generation inconsistency via uncertainty-aware learning," *arXiv:2411.18866*, 2024.

[14] Z. Zheng, J. Zhu, W. Ji, Y. Yang, and T.-S. Chua, "3d magic mirror: Clothing reconstruction from a single image via a causal perspective," *arXiv:2204.13096*, 2022.

[15] Y.-W. Chao, W. Yang, Y. Xiang, P. Molchanov, A. Handa, J. Tremblay, Y. S. Narang, K. Van Wyk, U. Iqbal, S. Birchfield, J. Kautz, and D. Fox, "Dexycb: A benchmark for capturing hand grasping of objects," in *CVPR*, 2021.

[16] U. Iqbal, P. Molchanov, T. B. J. Gall, and J. Kautz, "Hand pose estimation via latent 2.5D heatmap regression," in *ECCV*, 2018.

[17] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt, "Ganerated hands for real-time 3D hand tracking from monocular RGB," in *CVPR*, 2018.

[18] D. Tang, H. Jin Chang, A. Tejani, and T.-K. Kim, "Latent regression forest: Structured estimation of 3D articulated hand posture," in *CVPR*, 2014.

[19] C. Zimmermann and T. Brox, "Learning to estimate 3D hand pose from single RGB images," in *ICCV*, 2017.

[20] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: a skinned multi-person linear model," *ACM Trans. Graph.*, 2015.

[21] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM Trans. Graph.*, 2017.

[22] L. Yang, S. Li, D. Lee, and A. Yao, "Aligning latent spaces for 3d hand pose estimation," in *ICCV*, 2019.

[23] L. Yang and A. Yao, "Disentangling latent hands for image synthesis and pose estimation," in *CVPR*, 2019.

[24] P. Chen, Y. Chen, D. Yang, F. Wu, Q. Li, Q. Xia, and Y. Tan, "I2uv-handnet: Image-to-uv prediction network for accurate and high-fidelity 3d hand mesh modeling," in *ICCV*, 2021.

[25] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *CVPR*, 2021.

[26] S. Liu, W. Wu, J. Wu, and Y. Lin, "Spatial-temporal parallel transformer for arm-hand dynamic estimation," in *CVPR*, 2022.

[27] M. Li, L. An, H. Zhang, L. Wu, F. Chen, T. Yu, and Y. Liu, "Interacting attention graph for single image two-hand reconstruction," in *CVPR*, 2022.

[28] T. H. E. Tse, F. Mueller, Z. Shen, D. Tang, T. Beeler, M. Dou, Y. Zhang, S. Petrovic, H. J. Chang, J. Taylor *et al.*, "Spectral graphormer: Spectral graph-based transformer for egocentric two-hand reconstruction using multi-view color images," in *ICCV*, 2023.

[29] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee, "Perspective transformer nets: learning single-view 3d object reconstruction without 3d supervision," in *NeurIPS*, 2016.

[30] N. Siddharth, B. Paige, J.-W. van de Meent, A. Desmaison, N. D. Goodman, P. Kohli, F. Wood, and P. H. Torr, "Learning disentangled representations with semi-supervised deep generative models," in *NeurIPS*, 2017.

[31] J. Wu, Y. Wang, T. Xue, X. Sun, B. Freeman, and J. Tenenbaum, "Marrnet: 3d shape reconstruction via 2.5d sketches," in *NeurIPS*, 2017.

[32] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *CVPR*, 2017.

[33] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: deep hierarchical feature learning on point sets in a metric space," in *NeurIPS*, 2017.

[34] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *ICCV*, 2021.

[35] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, "Pct: Point cloud transformer," *Computational Visual Media*, 2021.

[36] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *CVPR*, 2019.

[37] Y. Yang, C. Feng, Y. Shen, and D. Tian, "Foldingnet: Point cloud auto-encoder via deep grid deformation," in *CVPR*, 2018.

[38] T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry, "A papier-mâché approach to learning 3d surface generation," in *CVPR*, 2018.

[39] C. Wang, D. Xu, Y. Zhu, R. Martin-Martin, C. Lu, L. Fei-Fei, and S. Savarese, "Densefusion: 6d object pose estimation by iterative dense fusion," in *CVPR*, 2019.

[40] W. Chen, X. Jia, H. J. Chang, J. Duan, L. Shen, and A. Leonardis, "Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism," in *CVPR*, 2021.

[41] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv:1711.00199*, 2017.

[42] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, "Normalized object coordinate space for category-level 6d object pose and size estimation," in *CVPR*, 2019.

[43] S. Ren, Y. Zeng, J. Hou, and X. Chen, "Corri2p: Deep image-to-point cloud registration via dense correspondence," *TCSVT*, 2023.

[44] J. Li and G. Hee Lee, "Deepi2p: Image-to-point cloud registration via deep classification," in *CVPR*, 2021.

[45] Z. Lin, C. Ding, H. Yao, Z. Kuang, and S. Huang, "Harmonious feature learning for interactive hand-object pose estimation," in *CVPR*, 2023.

[46] Z. Fan, M. Parelli, M. E. Kadoglou, X. Chen, M. Kocabas, M. J. Black, and O. Hilliges, "Hold: Category-agnostic 3d reconstruction of interacting hands and objects from video," in *CVPR*, 2024.

[47] Z. Chen, S. Chen, C. Schmid, and I. Laptev, "gSDF: Geometry-driven signed distance functions for 3d hand-object reconstruction," in *CVPR*, 2023.

[48] W. Dai, J. Li, D. LI, A. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," in *NeurIPS*, 2023.

[49] B. F. Labs, "Flux," https://github.com/black-forest-labs/flux, 2024.

[50] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[51] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, and C. Lu, "Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation," in *CVPR*, 2021.

[52] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "Shapenet: An information-rich 3d model repository," 2015.

[53] A. Miller and P. Allen, "Graspit! a versatile simulator for robotic grasping," *IEEE Robotics & Automation Magazine*, 2004.

[54] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv:1506.03365*, 2015.

[55] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[56] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *TPAMI*, 2021.

[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv*, 2014.
[59] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *ACM Trans. Graph.*, 2019.