IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 June 2024, accepted 11 August 2024, date of publication 19 August 2024, date of current version 13 September 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3445993*

RESEARCH ARTICLE

# ShapeGraFormer: GraFormer-Based Network for Hand-Object Reconstruction From a Single Depth Map

**AHMED TAWFIK ABOUKHADRA[1,2], JAMEEL MALIK[1,3], NADIA ROBERTINI[1], AHMED ELHAYEK[4], AND DIDIER STRICKER[1,2]**

[1]Augmented Vision Group, German Research Center for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany
[2]Department of Computer Science, University of Kaiserslautern-Landau (RPTU), 67663 Kaiserslautern, Germany
[3]School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan
[4]Artificial Intelligence Department, College of Computer and Cyber Sciences, University of Prince Mugrin (UPM), Madinah 42241, Saudi Arabia

Corresponding author: Ahmed Tawfik Aboukhadra (ahmed_tawfik.aboukhadra@dfki.de)

This work was supported in part by the Federal Ministry of Education and Research of the Federal Republic of Germany (BMBF) through DECODE and GreifbAR under Grant 01IW21001 and Grant 16SV8732.

**ABSTRACT** 3D reconstruction of hand-object manipulations is important for emulating human actions. Most methods dealing with challenging object manipulation scenarios focus on hands reconstruction in isolation, ignoring physical and kinematic constraints due to object contact. Some approaches produce more realistic results by jointly reconstructing 3D hand-object interactions. However, they focus on coarse pose estimation or rely upon known hand and object shapes. We propose an approach for realistic 3D hand-object shape and pose reconstruction from a single depth map. Unlike previous work, our voxel-based reconstruction network regresses the vertex coordinates of a hand and an object and reconstructs more realistic interaction. Our pipeline additionally predicts voxelized hand-object shapes, having a one-to-one mapping to the input voxelized depth. Thereafter, we exploit the graph nature of the hand and object shapes, by utilizing the recent GraFormer network with positional embedding to reconstruct shapes from template meshes. In addition, we show the impact of adding another GraFormer component that refines the reconstructed shapes based on the hand-object interactions and its ability to reconstruct more accurate object shapes. From those contributions, we name our system *ShapeGraFormer*. We perform an extensive evaluation on the HO-3D and DexYCB datasets and show that our method outperforms existing approaches in hand reconstruction and produces plausible reconstructions for the objects.

**INDEX TERMS** Computer vision, deep learning, graph convolutional network, hand-object 3D reconstruction, pose estimation, Transformers.

## I. INTRODUCTION

Understanding and reconstructing hand and object interactions in 3D is important for analyzing and imitating human behavior. Modeling hand-object interactions realistically has applications in several fields, including robotics, virtual reality, and augmented reality. The last decade has witnessed rapid advance in 3D hand pose [1], [2], [3], [4], [5], [6], [7], [8], [9] and object [10], [11], [12], [13], [14] estimation in isolation. In contrast, reconstructing a hand and an object simultaneously from a monocular image has received lesser attention. Besides the common issues from complex pose variation, clutter, and self-occlusion, methods for reconstructing hands and objects that are in close contact have to cope with *mutual occlusions*. Existing methods, dealing with challenging object manipulation scenarios, tend to focus on hand reconstruction alone [15], [16]. Recent approaches to jointly reconstruct hand and object, often neglect the intrinsic kinematic and physical correlation, which exists among the two [17], [18], [19], [20]. Approaches

The associate editor coordinating the review of this manuscript and approving it for publication was Davide Patti.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

exploiting that mutual relation typically focus on coarse pose estimation or assume known hand and object shapes [21], [22], [23], [24], [25], [26], [27].

In this paper, we propose one of the first approaches to jointly reconstruct physically valid hand and object shapes from a single depth map. In contrast to most methods, ours can generalize to different hand models and unknown object shapes, by directly regressing mesh vertices, rather than model parameters. We avoid perspective distortion and scale ambiguities, typical for RGB image-based methods by working exclusively in the 3D domain. The input of our deep network is a 3D voxelized grid of a given depth map, centered around the hand-object interaction. The output consists of (i) 3D heatmaps that describe the location of hand-object pose keypoints (ii) hand-object shape predictions in voxelized form, and (iii) the corresponding 3D hand-object mesh vertex coordinates.

To effectively tackle the problem of simultaneous hand-object pose and shape reconstruction, we propose a novel architecture based on a Graph Convolutional network and Multi-headed Attention layers. Specifically, we introduce the following novel modules:

1) *PoseNet* & *VoxelNet*: Two 3D-to-3D voxel-based networks for hand-object pose and shape estimation, respectively;
2) *ShapeGraFormer*: State-of-the-art GraFormer (Transformers with Graph Convolutional layers) for Hand-Object shape reconstruction;
3) Positional Embedding layer based on the template meshes for the hand and the object;
4) Topologically consistent object mesh registration for optimal object modeling and shape prediction;

We validate our design choices and evaluate our approach both quantitatively and qualitatively. Our approach outperforms previous work on popular datasets [28], [29], as also reported on the challenge website,[1] with a minimal shape reconstruction improvement of 0.43 cm over the state-of-the-art.

## II. RELATED WORK

In this section, we discuss the existing methods for joint hand-object reconstruction from challenging monocular object manipulation scenarios. For a survey of works focusing on the reconstruction of hands and objects in isolation, please refer respectively to [9] and [11].

Most methods that jointly reconstruct 3D hand-object from monocular take single RGB [20], [23], [24], [26], [27], [30], [31] or RGB + D input [17], [21], [22]. Very few recent approaches consider single depth input [32], [33], [34], due to its intrinsic challenges. Nevertheless, the availability of depth information is a key factor in allowing proper, in-scale, and scene-dependent 3D shape reconstruction, required for e.g., virtual and augmented reality applications, especially for single-frame, one-shot approaches. While RGB-based approaches can rely on large labeled data for training their models [18], [23], [35], typically based on the MANO parametric hand model (hand Model with Articulated and Non-rigid defOrmations) [36], methods that rely on the depth channel have access to only a limited amount of data, because labeling of real scenes in the 3D domain are impractical. For this reason, most depth as well as some RGB + D approaches build their synthetic datasets in an attempt to improve their results [17], [23], [32], [33], [34]. Recently, some datasets have been introduced to bridge the gap between RGB and depth data availability, i.e., HO3D and DexYCB [28], [29]. Still, they only provide limited sample variation, especially in terms of number of objects considered. As a result, modeling and reconstructing the object of interaction remains an underconstrained, challenging problem. Many methods restrict themselves to objects with known shapes [21], [22] or reconstruct them on the fly, under certain object shape and visibility assumptions [34], [37], [38]. The remaining approaches only output coarse object pose, e.g., represented with bounding boxes [19], [24], [30], [31]. Some more sophisticated methods, define a deformable object model, typically based on a 3D sphere, capable of coarsely adapting to virtually any convex shape [20], [23], [37]. The resulting object reconstruction typically lacks surface details, due to over-smoothing. Nevertheless, we believe it to be a strong base with sufficient surface information to reliably reconstruct interactions with the hand surface. The remaining hand reconstruction as a standalone problem has been largely studied in the past [39]. However, modeling and reconstructing 3D hand-object interactions is still very challenging, especially in a monocular setting due to the large mutual occlusions. To simplify the task, many approaches focus on fewer model parameter estimation [23], [26]. In contrast, we directly regress hand-object vertices, which makes our method capable to generalize to different models or geometries. While many approaches independently reconstruct hand and object before putting them in context [23], [26], our approach is designed to simultaneously regress hand-object geometries, thus allowing us to implicitly study the underlying physical and kinematic correlation existing among the two. Another advantage of direct vertex regression is the possibility of reconstructing extra, more realistic deformation, which cannot be synthesized via model-parameter tuning. In this case, to avoid distortions in the 3D reconstructions, though, particular care has to be given to the algorithm design, especially to the shape estimation components. To avoid perspective distortion from the start, we convert the input depth map to a 3D point cloud and base the remaining algorithmic steps on the corresponding voxelized domain, following the approach from Malik et al. [9] (HandVoxNet). The core of our pipeline is based on a Graph Convolutional Network (GCN), which has been shown to effectively tackle shape reconstruction problems on graph-structured data, such as mesh topology [40]. In contrast to the RGB-based approach by Aboukhadra et al. [20] (THOR-Net), our depth-based

[1] https://codalab.lisn.upsaclay.fr/competitions/4393#results

method reconstructs hand-object geometries in one shot, thus significantly reducing computational costs and training time. We qualitatively and quantitatively demonstrate the effectiveness of our *ShapeGraFormer* component, comprising a combination of GCN and Multi-headed Attention layers, as in [41], in the simultaneous reconstruction of hand and object interaction as well as in shape refinement. For the latter, we demonstrate its effectiveness in improving physical hand-object interactions, without the explicit need for expensive physical simulation [21] or penetration and contact loss, as required in previous work [23], [26].

## III. METHOD

We design a voxel-based 3D CNN along with a *ShapeGraFormer* network to reconstruct plausible 3D hand-object shapes in a single forward pass from an input depth image. Our pipeline is depicted in Fig 1. In a preprocessing step, we convert the input depth map to its voxelized form $V_D$ by projecting the raw depth image pixels into a cubic binary 3D grid around the hand-object interaction space, similar to [9]. Given $V_D$, the first network component in the pipeline, *PoseNet*, predicts 3D hand and object pose in the form of 3D heatmaps, resp. $\hat{P}^H$ and $\hat{P}^O$, see Section III-B. The resulting heatmaps concatenated with $V_D$ are forwarded to the second network component, *VoxelNet*, which produces a voxelized shape representation of the hand and the object, resp. $\hat{V}^H$ and $\hat{V}^O$, see Section III-C. The voxelized depth $V_D$ along with the intermediate voxelized representation and the features of the *VoxelNet* serve as input to the next network component, *ShapeGraFormer*, which regresses topologically consistent hand and object vertices, see Section III-D. We describe our hand and object models in the next Section.

### A. HAND AND OBJECT MODELS

#### 1) HAND MODEL

To represent hands, we use the MANO parametric hand model [36], which maps joint angles and shape parameters to a triangulated mesh representation $S^H := \{v^H, f^H\}$, consisting of $|v_H| = 778$ 3D vertices, and a set of 3D hand-skeleton joints $\mathcal{J}$, where $|\mathcal{J}| = 21$, representing the hand pose.

#### 2) OBJECT MODEL

Since objects differ greatly in terms of shape and size, direct regression of object vertices from a set of topologically inconsistent meshes results in strong noise in the output. In order to bring all the known object geometries into the same topologically consistent representation $S^O := \{v^O, f^O\}$, we deform a source mesh via a set of vertex displacements $D^j := \{d_i^j, \forall v_i^O\}$, scale $c_j$ and translation $t_j$ to approximate all target objects $S^j$ as:

$$S^j \sim c^j(S^O + D^j) + t^j, \forall j \tag{1}$$

Our source mesh is a sphere obtained performing 4 subdivisions of a unit, i.e.,radius equals 1, regular icosahedron centered at the origin, each subdivision generating four new faces per face, resulting in a total of $|v^O| = 2562$ vertices and $|f^O| = 5120$ faces. The object pose is defined as the 3D bounding box, consisting of eight 3D corner coordinates.

#### 3) OBJECT APPROXIMATION PROCEDURE

To approximate differently shaped objects, we first scale ($c^j$) and translate ($t^j$) the target mesh $S^j$ to fit inside the sphere. Then, to learn the set of displacements $D^j$, we minimize the chamfer distance $E^C$ between the predicted and target mesh computed on a total of $5,000$ surface samples. We additionally enforce surface smoothness by adding the following shape regularizers to the objective: (i) surface laplacian smoothness $E^L$ which ensures smooth vertex positions between neighboring vertices, (ii) normals consistency $E^N$ across neighboring faces, and (iii) edge length consistency $E^E$ across the entire deformed mesh which minimizes the length of edges. We minimize the weighted summation of all mentioned terms in the following equation using stochastic gradient descent (SGD). The full implementation of spherical object approximation can be found in the PyTorch3D library.[2]

$$E = w_C E^C + w_L E^L + w_N E^N + w_E E^E \tag{2}$$

### B. PoseNet: 3D POSE ESTIMATION

The first component of our network pipeline, *PoseNet*, simultaneously estimates 3D hand joint location and 3D bounding box corners of the object from an input voxelized depth map. We modify the V2V-PoseNet architecture of [42] by introducing a new $1 \times 1 \times 1$ volumetric convolutional back-layer, targeted to predict 3D object pose in parallel to the 3D hand pose. The input depth map is converted to a 3D binary voxelized representation $V_D$ of size $88 \times 88 \times 88$. Each voxel value is set to 1 when occupied and 0 otherwise, i.e.,$V_D \in \{0, 1\}$. The cubic size is fixed to the empirically found value of 200 mm. The output of the network consists of a set of 3D heatmaps: (i) One set for each $j$ hand joints coordinates $\hat{P}_j^H(v)$, and (ii) one for each $b$ object bounding box corners coordinates $\hat{P}_b^O(v)$. All heatmaps are discretized on a grid of size $44 \times 44 \times 44$. Ground-truth heatmaps $P_j^H(v)$ and $P_b^O(v)$ are generated by applying a 3D Gaussian centered on the ground-truth locations with fixed standard deviation. The training loss between the predicted and targeted heatmaps is calculated by mean-squared error (MSE):

$$\mathcal{L}_{pose} = \sum_v \sum_{j=1}^{|J|} ||\hat{P}_j^H - P_j^H|| + \sum_{b=1}^{|B|} ||\hat{P}_b^O - P_b^O|| \tag{3}$$

where $J = 21$ is the number of hand joints and $B = 8$ is the number of object corners In the above formula, we omitted all dependencies on the voxels $v$ for brevity.

[2]https://pytorch3d.org/tutorials/deform_source_mesh_to_target_mesh

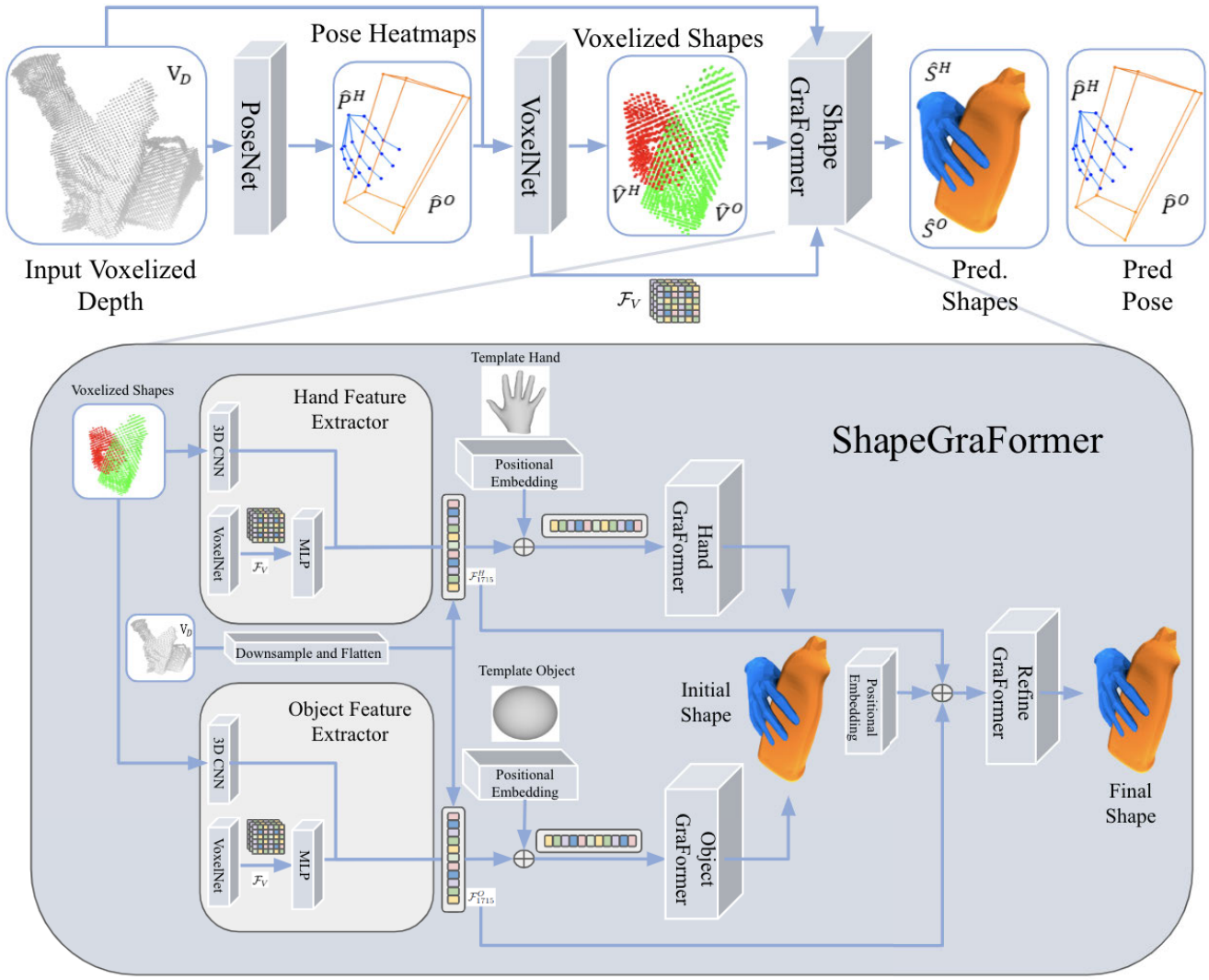



**FIGURE 1.** Overview of our pipeline. Our method takes as input a single depth image that is converted to a 3D voxelized representation and outputs 3D realistic hand-object interactions. The input depth is first forwarded through a sequence of three components, namely *PoseNet*, *VoxelNet*, and *ShapeGraFormer*, predicting respectively hand-object pose heatmaps, voxelized shape representations, and topologically-consistent shapes.

### C. VoxeLNet: 3D VOXELIZED SHAPE

Given $V_D$, $\hat{P}^H$ and $\hat{P}^O$, the second component of our pipeline, *VoxelNet*, predicts hand-object voxelized shapes $\hat{V}^H(v)$ and $\hat{V}^O(v)$ for all existing voxels $v$ in the grid. The 3D CNN-based architecture of *VoxelNet* is inspired by the work of Malik et al. [9]. We introduce an additional 3D convolutional layer to predict the voxelized object shape. Our voxelized hand-object shapes are defined in the range $[0, 1]$, as done in the previous work. The predicted voxelized shapes represent complete surface representation for both the hand and the object, including both visible and occluded surface information, as learned from the dataset. Thus, entailing richer information for the next algorithmic steps.

To train *VoxelNet*, we use the per-voxel combined sigmoid activation function with binary cross-entropy loss for the voxelized hand shape (similarly for the voxelized object shape):

$$\mathcal{L}^H_{voxel}(v) = -(V^H log(\hat{V}^H) + (1 - V^H)log(1 - \hat{V}^H)) \quad (4)$$

where $V^H$ and $\hat{V}^H$ are the ground truth and the estimated voxelized hand and object shapes, respectively.

### D. SHAPEGRAFORMER

To obtain topologically-coherent hand-object mesh representation, from the voxelized hand and object predictions obtained from *VoxelNet*, we train a new *ShapeGraFormer* [41]. The *ShapeGraFormer* includes Graph Convolutional layers and Multi-headed Attention layers to convert depth features into a pose and a shape for both the hand and the object. Although it was originally designed to lift 2D poses to 3D, GraFormer's design combines the advantages of Graph Convolutional Networks and Transformers making it capable of effectively solving any problem that can be represented

**FIGURE 2.** The complete collection of our sphere-based object approximations used as ground-truth.

as a graph. Both the hand and the spherical objects have consistent topology and hence can be represented as a graph. Therefore, in our method, we utilize three separate GraFormers, namely: hand GraFormer, object GraFormer, and refinement GraFormer, with two feature extractors: a hand feature extractor and an object feature extractor as described in the following subsection.

Specifically, the network outputs hand-object vertex coordinates $v^H$ and $v^O$, respectively from the MANO hand model $S^H$ and the sphere-based object approximated shape $S^O$. The hand and object vertices loss for training is defined using MSE:

$$\mathcal{L}_{shape} = \frac{\sum_i (v_i^H - \hat{v}_i^H)^2}{|v^H|} + \frac{\sum_j (v_j^O - \hat{v}_j^O)^2}{|v^O|} \tag{5}$$

where $\hat{v}_i^H$ and $v_i^H$ are respectively the predicted and ground-truth $i$-th hand vertex coordinates, and $\hat{v}_j^O$ and $v_j^O$ are predicted and ground-truth object vertex coordinates.

### 1) FEATURE EXTRACTOR AND GRAPH INITIALIZATION

For the graph initialization, we use the outputs of the *VoxelNet* and the raw voxelized depth as shown in Fig. 1. Namely, we extract feature maps $\mathcal{F}_V$ from the *VoxelNet* and pass them through an MLP that converts them into a feature vector of size 256. Additionally, we utilize a simple 3D CNN to extract features from the voxelized shapes and reduce them to 128 features. Furthermore, a 3D Max Pool layer reduces the size of the raw voxelized depth from $44 \times 44 \times 44$ to $11 \times 11 \times 11$ creating a 1331 feature vector. These features are then combined to create a hand feature vector $\mathcal{F}_{1715}^H$ of size 1715 to initialize the graph vertices as shown in Fig. 1. The same operation is repeated for the object's vertices resulting in $\mathcal{F}_{1715}^O$. To initialize the adjacency matrix of the graph

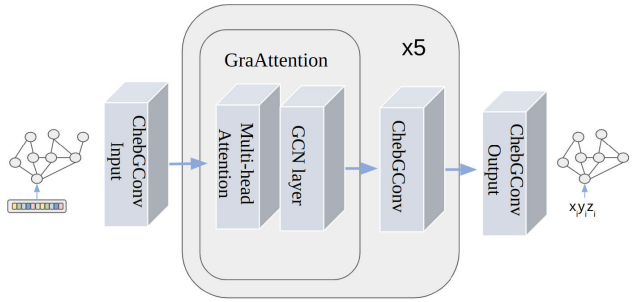



**FIGURE 3.** Illustration of the GraFormer architecture.

layers, we use the mesh faces of the MANO model along with the faces of the spherical mesh as described in Section III-A.

### 2) POSITIONAL EMBEDDING

In order to generate distinct features for each vertex in the graph, we propose a positional embedding layer that converts the vertices of the template meshes for both the hand and object into positional vectors $\mathcal{E}_i^p$ of the same size as the feature vector where $i$ is the index of the $i$th vertex in the combined hand-object graph. The positional vectors $\mathcal{E}_i^p$ are then accumulated with the shared feature vectors $\mathcal{F}_{1715}^H$ and $\mathcal{F}_{1715}^O$ depending on whether the $i$th vertex belongs to the hand or the object in order to create a unique representation for each vertex. For the hand mesh, we adopt the default MANO hand as the template, while for the object mesh, we use the sphere that is utilized for deforming objects, see Section III-A.

### 3) GRAFORMER DETAILS

Each GraFormer consists of five consecutive GraAttention components followed by Chebyshev Graph Convolutional layers as described in [41] and shown in Fig. 3. A GraAttention component consists of a Multi-headed Attention layer where the hidden dimension is 128 and the number of heads is 4, following the ablation study mentioned in [41] where they studied the impact of hidden dimension size and number of layers in the GraFormer on pose estimation. Compared to normal Transformers, the last fully connected layer of the Multi-headed Attention in the GraFormer is a graph convolutional layer, not a feed-forward layer. The GraFormer also contains an input layer that maps the feature vector to the hidden dimension size and an output layer that maps the hidden dimension into the corresponding 3D coordinate value for each vertex in the graph.

### 4) REFINEMENT GRAFORMER

To enhance the realistic appearance and correct minor shape-related artifacts in the predicted object shapes, we add another GraFormer for refinement. We utilize the initial shape produced by the Hand and Object GraFormer as input to another positional embedding layer while we use $\mathcal{F}_{1715}^H$ and $\mathcal{F}_{1715}^O$ to initialize the new graph. In the case of using a refinement GraFormer during training, we get 2 separate

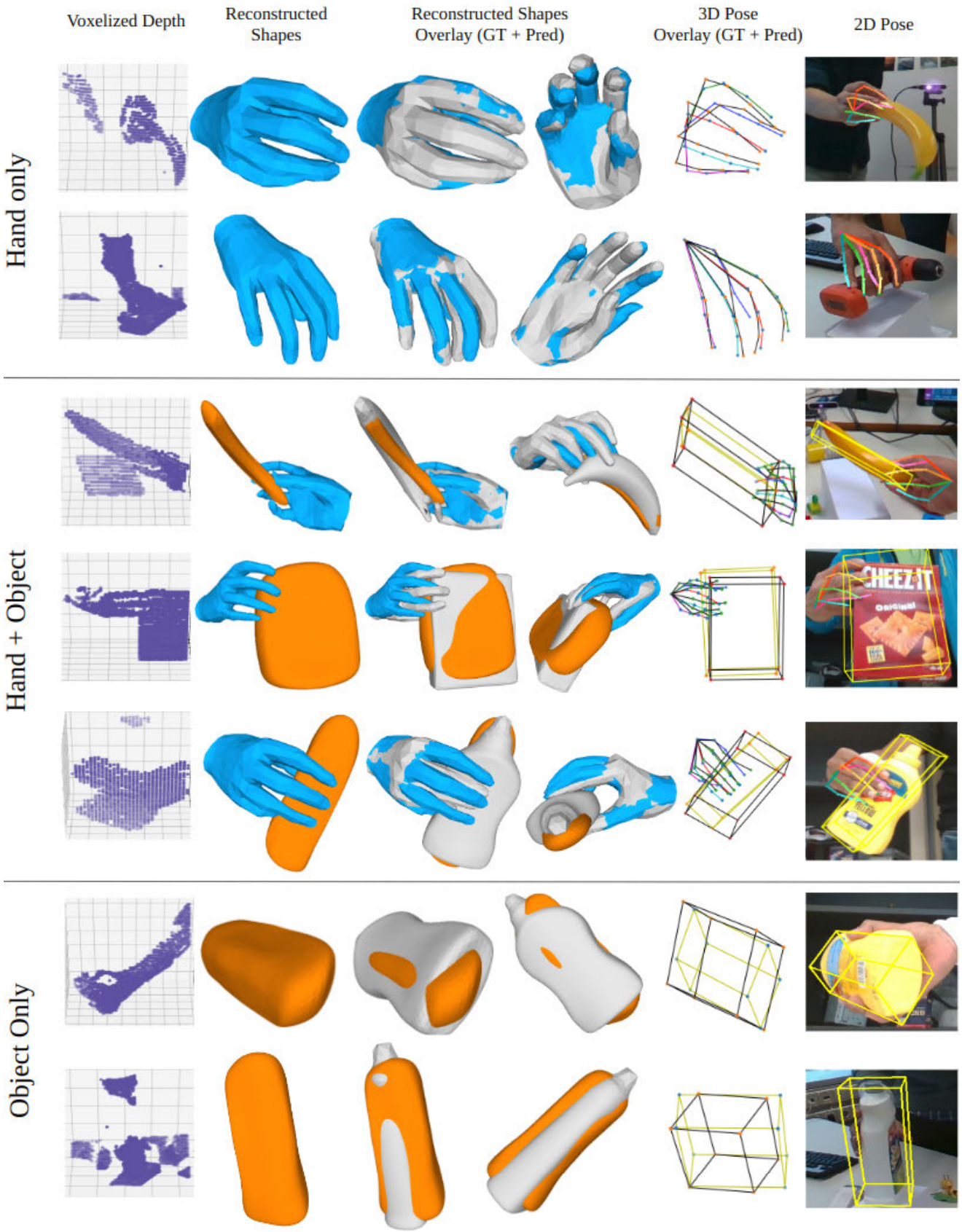



**FIGURE 4.** Reconstruction results at different steps. Each row shows a different sample from $\mathcal{D}^{valid}$ or $\mathcal{D}^{eval}$ from HO-3D (v3).

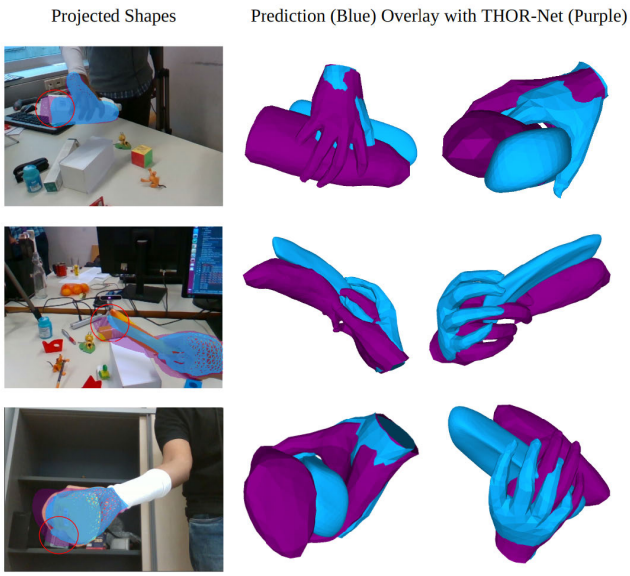



**FIGURE 5. Comparison of our hand-object reconstruction with the THOR-Net pipeline on the evaluation set $\mathcal{D}^{eval}$. Our hand-object reconstruction approach is shown to generate more accurate hand and object shapes.**

shapes from the *ShapeGraFormer* and apply the same loss function mentioned in III-D on both of them.

### E. TRAINING AND IMPLEMENTATION DETAILS

We train all the components of our network on fully annotated public hand-object pose and shape datasets, namely, HO-3D [28] and DexYCB [29]. For training, we use the Adam [43] optimizer with a learning rate set to .001. For improved convergence, we train *PoseNet* separately, then fix the obtained weights to train the remaining network components. All learning and inference are implemented in PyTorch and are conducted on an NVIDIA A100 GPU.

## IV. EXPERIMENTS

In this section, we quantitatively and qualitatively evaluate the effectiveness of our 3D hand-object reconstruction pipeline on two popular datasets, namely HO-3D dataset (version v2 and v3) [28] and DexYCB [29]. We additionally compare our pipeline with state-of-the-art approaches in Section IV-B. For quantitative evaluation and comparisons, we use the following two metrics: (i) the average 3D joint location error ($\mathcal{J}$ err.), and (ii) the average 3D vertex location error ($\mathcal{V}$ err.) over all test frames.

### A. DATASETS

#### 1) HO-3D

The HO-3D dataset [28], is a publicly available dataset with 3D pose annotations for hands interacting with objects captured from third-person views. The dataset has multiple versions and we report the results on v2 and v3 of the dataset. For HO-3D (v3), the training set $\mathcal{D} := \{\mathcal{D}^{train}, \mathcal{D}^{valid}\}$ contains annotations for $|\mathcal{D}^{train}| = 71,662$ and $|\mathcal{D}^{valid}| = 10,927$ images and a total of 55 sequences and 10 different objects from the YCB-Video dataset [44], 9 for the training set and 1 (unseen) for the evaluation set. The evaluation set $\mathcal{D}^{eval}$ comprises 13 sequences with a total of $|\mathcal{D}^{eval}| = 20,137$ frames addressing challenging scenarios: namely, (i) 3 sequences with (2) seen objects and seen hands, (ii) 5 sequences with (1) seen object but unseen hands, and (iii) 5 sequences with seen hands but 1 unseen object. Hands in the evaluation set are just annotated with the wrist coordinates, while the full hand is not annotated. Object pose and shape are annotated over all available sets.

#### 2) DexYCB

To extend our evaluation, we also train our network on the DexYCB dataset [29]. DexYCB contains hand pose and shape and 6D object pose annotations for $582k$ frames recorded on 10 different subjects, using 20 different objects and 8 views. We use the S1 evaluation setup as specified by the authors where $\mathcal{D}^{valid}$ contains 1 unseen subject and $\mathcal{D}^{eval}$ contains 2 unseen subjects and $\mathcal{D}^{train}$ contains 7 subjects and all different objects. The exact split sizes are: $|\mathcal{D}^{train}| = 407,088$, $|\mathcal{D}^{valid}| = 58,592$, and $|\mathcal{D}^{eval}| = 116,288$.

In the next section, we show results on $\mathcal{D}^{eval}$ for both datasets.

### B. EVALUATION

In this section, we evaluate the hand-object shape and pose reconstruction and compare it to state-of-the-art approaches in challenging scenarios.

#### 1) METHODS FOR COMPARISON

We compare our work on HO-3D quantitatively with six different methods. The work of Hasson et al. [23] and THOR-Net [20] are the most related to ours in terms of the goals. However, as they focus on RGB inputs, comparisons are made up to scale. Malik et al. [9] on the other hand, is based on depth inputs and has a comparable voxel-based network pipeline to our design. However, they ignore the presence of an object and reconstruct hands in isolation. We include the representative RGB-based hand reconstruction approach of Hampali et al. [28], and ArtiBoost [31] for completeness. For a fair comparison, all methods have been (re-)trained on the HO-3D dataset. The first two methods provide publicly available results (hand-only), which we report in Table 1. We re-implemented HandVoxNet following the authors' instructions and trained all the network components on HO-3D. In addition to that, we also report the root-relative pose estimation error in Table 2 and compare it to two benchmark methods mentioned by the DexYCB authors. We also show qualitative samples for hand reconstruction from DexYCB in Fig. 6.

#### 2) HAND RECONSTRUCTION

Quantitative evaluation on hand pose and shape reconstruction, shows that our method outperforms existing methods,

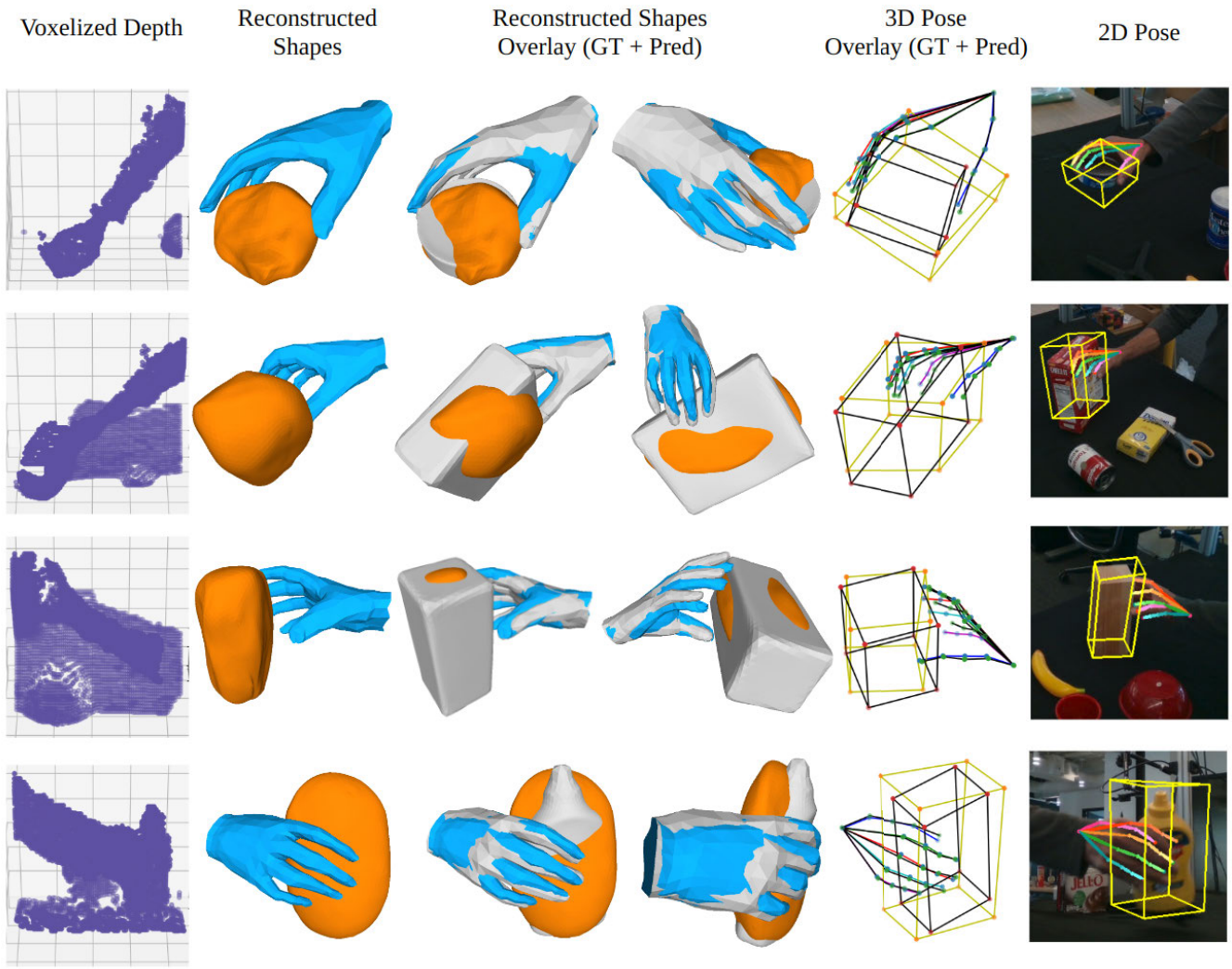



**FIGURE 6.** Reconstruction results at different steps. Each row shows a different sample from $\mathcal{D}^{eval}$ from DexYCB.

**TABLE 1.** Quantitative results on the evaluation set $\mathcal{D}^{eval}$ of HO-3D [28] dataset. Upper table: HO-3D (v2), lower table: HO-3D (v3).

| Method | $\mathcal{J}$ err. (*cm*) | $\mathcal{V}$ err. (*cm*) |
|---|---|---|
| Hampali *et al.* [28] | 8.42 | 8.34 |
| Hasson *et al.* [23] | 4.97 | 4.96 |
| HandVoxNet [9] | 2.77 | 2.67 |
| HandVoxNet++ [40] | 2.46 | 2.70 |
| THOR-Net [20] | 2.63 | 2.63 |
| Ours (hand-object, w/ contact loss) | 2.27 | 2.20 |
| **Ours (hand-only, w/o refinement)** | **2.13** | **2.07** |
| ArtiBoost [31] | 2.26 | 10.95 |
| THOR-Net [20] | 2.56 | 2.37 |
| Ours (hand-object, w/ refinement) | 2.15 | 2.09 |
| Ours (hand-only, w/ refinement) | 2.00 | 1.94 |
| **Ours (hand-only, w/o refinement)** | **1.99** | **1.94** |

see Table 1 and Figures 4, 5. We achieve a minimum pose and shape prediction improvement respectively of 0.27 *cm* and 0.43 *cm* over the state-of-the-art. The experiments show that training the network for hand reconstruction alone without object reconstruction improves hand reconstruction and that the refinement stage has no impact on hand reconstruction as shown in Table 1. However, joint training for the hand and the object outperforms other methods. Fig. 5 shows our results on a few frames of the challenging evaluation sequence compared to THOR-Net. In comparison to THOR-Net, our approach reconstructs more accurate hand shapes, as it better exploits hand-object kinematic correlation and depth information. These are implicitly learned while predicting both interacting shapes simultaneously. In Table 2, we show an improvement of 0.53 *cm* in hand pose estimation compared to benchmark results on the DexYCB dataset. Fig. 6 also shows the qualitative reconstruction results on DexYCB without the additional refinement network.

### 3) OBJECT RECONSTRUCTION

We found that topological consistency is the key factor allowing *ShapeGraFormer* to predict smooth vertices point clouds across all sequences, without the need for additional smoothness constraints, see Fig. 4. After sphere-based

**TABLE 2.** Root-relative hand pose estimation results on $\mathcal{D}^{eval}$ of DexYCB [29] dataset (S1) in comparison to benchmark results.

| Method | $\mathcal{J}$ err. (*cm*) | $\mathcal{V}$ err. (*cm*) |
|---|---|---|
| A2J [45] | 2.55 | - |
| Spurr *et al.* [46] | 2.27 | - |
| **Ours** | **1.74** | **2.65** |

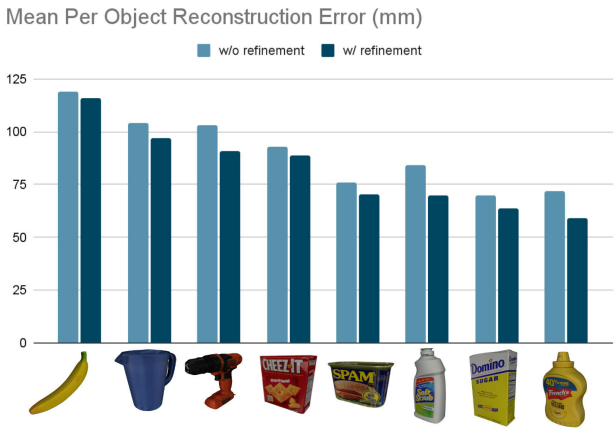



**FIGURE 7.** Object reconstruction error ($\mathcal{V}$ err. in *mm*) on different objects from $\mathcal{D}^{valid}$ and $\mathcal{D}^{eval}$ of HO-3D. The results show that adding a refinement GraFormer improves object reconstruction.

registration, all objects share the same topology and number of vertices. We believe our choice for the sphere resolution to be a good trade-off between approximation quality and the number of vertices. Sphere-based approximation implicitly repairs irregularities in the target object shapes, making it best suited for hand-object prediction, at the cost of over-smoothed sharp edges and tiny surface details, see Fig. 2.

We evaluate object reconstruction on $\mathcal{D}^{valid}$ and $\mathcal{D}^{eval}$ of HO-3D dataset, see Fig. 7 and Table 3, as well as on the DexYCB dataset, see Fig. 6. The additional hand-object refinement step improved object reconstruction error on all objects. This suggests that the refinement network utilizes hand information to improve object reconstruction. We notice that even in the presence of inaccurate pose prediction as input, our approach recovers smooth object shapes, see Fig. 4. Compared to THOR-Net our approach tends to oversmooth object edges, see Fig. 5, possibly due to the simplified reconstruction approach.

**TABLE 3.** Object's reconstruction error ($\mathcal{V}$ err. (*cm*)) on a selected set of objects from HO-3D and DexYCB.

| | Bottle | Box | Can | Marker | Wood |
|---|---|---|---|---|---|
| $\mathcal{V}$ err. (*cm*) | 5.9 | 6.4 | 7.0 | 9.7 | 11.5 |

### C. ABLATION STUDY

To validate our design choices for the *ShapeGraFormer*, we conduct an ablation study to study the modality of $\mathcal{F}^H_{1715}$ and the choice of layers within the GraFormer and study the impact on hand reconstruction. The ablation study in Table 4 shows that combining voxelized depth $V_D$, $\mathcal{F}_V$, and $\hat{V}^H$ as input to the *ShapeGraFormer* yields the best results. Furthermore, the mixture of graph layers with transformers in the design of the GraFormer is critical to achieving the best performance.

**TABLE 4.** Top: An ablation study on the modality of $\mathcal{F}^H_{1715}$. Bottom: An ablation study on the GraFormer design choice.

| Experiment | $\mathcal{J}$ err. (*cm*) | $\mathcal{V}$ err. (*cm*) |
|---|---|---|
| ShapeGraFormer (w/o $V_D$) | 2.01 | 2.06 |
| ShapeGraFormer (w/o $\mathcal{F}_V$) | 2.05 | 2.01 |
| ShapeGraFormer (w/o $\hat{V}^H$) | 2.00 | 1.95 |
| **ShapeGraFormer** (w/ $V_D \oplus \mathcal{F}_V \oplus \hat{V}^H$) | **1.99** | **1.94** |
| ShapeGraFormer (w/o GCN) | 19.32 | 19.36 |
| ShapeGraFormer (w/o Transformer) | 2.06 | 2.01 |
| **ShapeGraFormer** (w/ GCN + Transformer) | **1.99** | **1.94** |

#### 1) CONTACT LOSS: PENETRATION AVOIDANCE AND CONTACT ENFORCEMENT

We test the impact of differentiable contact loss, consisting of an attraction $\mathcal{L}_{attraction}$ and a repulsion $\mathcal{L}_{repulsion}$ term, similar to Hasson et al. [23]. $\mathcal{L}_{attraction}$ is aimed at enforcing hand-object contact, by penalizing the distance between the object and the fingertips, while $\mathcal{L}_{repulsion}$ penalizes mesh interpenetration. As shown in Table 1, the additional loss term does not introduce a tangible increase in performance. Thus, suggesting our independent Refinement GraFormer component implicitly and successfully learns valid hand-object interactions.

## V. CONCLUSION

In this paper, we propose one of the first methods for realistic hand-object pose and shape reconstruction from a single depth map. We introduce a novel 3D voxel-based GraFormer network pipeline, which reconstructs detailed 3D shapes via direct regression of mesh vertices. We conduct an ablation study to show the effectiveness of our design choices and the impact of utilizing the power of GCNs along with Transformers for hand-object shape estimation and refinement. We perform quantitative and qualitative analysis on the HO-3D dataset [28] and DexYCB dataset [29] and show outstanding comparative results with the state-of-the-art. In future work, we plan to address limitations such as inaccurate annotations. In addition, we will study RGB + D methods to utilize the extra features found in RGB frames.

## REFERENCES

[1] J. Malik, A. Elhayek, F. Nunnari, K. Varanasi, K. Tamaddon, A. Heloir, and D. Stricker, "DeepHPS: End-to-end estimation of 3D hand pose and shape by learning from synthetic depth," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2018, pp. 110–119.

[2] J. Malik, A. Elhayek, and D. Stricker, "Structure-aware 3D hand pose regression from a single depth image," in *Proc. Int. Conf. Virtual Reality Augmented Reality (EuroVR)*, 2018, pp. 3–17.

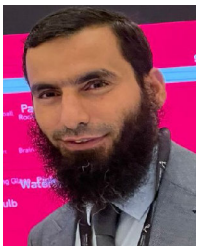
[3] L. Ge, Z. Ren, Y. Li, Z. Xue, Y. Wang, J. Cai, and J. Yuan, "3D hand shape and pose estimation from a single RGB image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10825–10834.

[4] D. Kulon, R. A. Güler, I. Kokkinos, M. M. Bronstein, and S. Zafeiriou, "Weakly-supervised mesh-convolutional hand reconstruction in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4989–4999.
[5] X. Zhang, Q. Li, H. Mo, W. Zhang, and W. Zheng, "End-to-end hand mesh recovery from a monocular RGB image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2354–2364.
[6] J. Yang, H. J. Chang, S. Lee, and N. Kwak, "SeqHAND: RGB-sequence-based 3D hand pose and shape estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 122–139.
[7] J. Malik, A. Elhayek, and D. Stricker, "WHSP-Net: A weakly-supervised approach for 3D hand shape and pose recovery from a single depth image," *Sensors*, vol. 19, no. 17, p. 3784, Aug. 2019.
[8] Y. Zhou, M. Habermann, W. Xu, I. Habibie, C. Theobalt, and F. Xu, "Monocular real-time hand shape and motion capture using multi-modal data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5345–5354.
[9] J. Malik, I. Abdelaziz, A. Elhayek, S. Shimada, S. A. Ali, V. Golyanik, C. Theobalt, and D. Stricker, "HandVoxNet: Deep voxel-based network for 3D hand shape and pose estimation from a single depth map," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7111–7120.
[10] D. Tzionas and J. Gall, "3D object reconstruction from hand-object interactions," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 729–737.
[11] B. Tekin, S. N. Sinha, and P. Fua, "Real-time seamless single shot 6D object pose prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 292–301.
[12] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2637–2646.
[13] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "PVNet: Pixel-wise voting network for 6DoF pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4556–4565.
[14] C. Wang, R. Martín-Martín, D. Xu, J. Lv, C. Lu, L. Fei-Fei, S. Savarese, and Y. Zhu, "6-PACK: Category-level 6D pose tracker with anchor-based keypoints," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 10059–10066.
[15] F. Mueller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, and C. Theobalt, "Real-time hand tracking under occlusion from an egocentric RGB-D sensor," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1163–1172.
[16] S. Baek, K. I. Kim, and T.-K. Kim, "Weakly-supervised domain adaptation via GAN and mesh model for estimating 3D hand poses interacting objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6120–6130.
[17] S. Sridhar, F. Mueller, M. Zollhöfer, D. Casas, A. Oulasvirta, and C. Theobalt, "Real-time joint tracking of a hand manipulating an object from RGB-D input," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 294–310.
[18] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 49–59.
[19] B. Doosti, S. Naha, M. Mirbagheri, and D. J. Crandall, "HOPE-Net: A graph-based model for hand-object pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6607–6616.
[20] A. T. Aboukhadra, J. Malik, A. Elhayek, N. Robertini, and D. Stricker, "THOR-Net: End-to-end graformer-based realistic two hands and object reconstruction with self-supervision," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 1001–1010.
[21] D. Tzionas, L. Ballan, A. Srikantha, P. Aponte, M. Pollefeys, and J. Gall, "Capturing hands in action using discriminative salient points and physics simulation," *Int. J. Comput. Vis.*, vol. 118, no. 2, pp. 172–193, Jun. 2016.
[22] A. Tsoli and A. A. Argyros, "Joint 3D tracking of a deformable object in interaction with a hand," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 484–500.
[23] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11799–11808.
[24] B. Tekin, F. Bogo, and M. Pollefeys, "H+O: Unified egocentric recognition of 3D hand-object poses and interactions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4506–4515.
[25] H. Zhang, Z.-H. Bo, J.-H. Yong, and F. Xu, "InteractionFusion: Real-time reconstruction of hand poses and deformable objects in hand-object interactions," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–11, Aug. 2019.
[26] Z. Cao, I. Radosavovic, A. Kanazawa, and J. Malik, "Reconstructing hand-object interactions in the wild," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12397–12406.
[27] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 568–577.
[28] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit, "HOnnotate: A method for 3D annotation of hand and object poses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3193–3203.
[29] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9040–9049.
[30] S. Hampali, S. D. Sarkar, M. Rad, and V. Lepetit, "Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3D pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11080–11090.
[31] L. Yang, K. Li, X. Zhan, J. Lv, W. Xu, J. Li, and C. Lu, "ArtiBoost: Boosting articulated 3D hand-object pose estimation via online exploration and synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2740–2750.
[32] M. Oberweger, P. Wohlhart, and V. Lepetit, "Generalized feedback loop for joint hand-object pose estimation," 2019, *arXiv:1903.10883*.
[33] J. Chen, M. Yan, J. Zhang, Y. Xu, X. Li, Y. Weng, L. Yi, S. Song, and H. Wang, "Tracking and reconstructing hand object interactions from point cloud sequences in the wild," 2022, *arXiv:2209.12009*.
[34] H. Zhang, Y. Zhou, Y. Tian, J.-H. Yong, and F. Xu, "Single depth view based real-time reconstruction of hand-object interactions," *ACM Trans. Graph.*, vol. 40, no. 3, pp. 1–12, Jun. 2021.
[35] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim, "First-person hand action benchmark with RGB-D videos and 3D hand pose annotations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 409–419.
[36] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM Trans. Graph.*, vol. 36, no. 6, p. 245, 2017.
[37] T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry, "A papier-mache approach to learning 3D surface generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 216–224.
[38] H. Kato, Y. Ushiku, and T. Harada, "Neural 3D mesh renderer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3907–3916.
[39] L. Huang, B. Zhang, Z. Guo, Y. Xiao, Z. Cao, and J. Yuan, "Survey on depth and RGB image-based 3D hand shape and pose estimation," *Virtual Reality Intell. Hardw.*, vol. 3, no. 3, pp. 207–234, Jun. 2021.
[40] J. Malik, S. Shimada, A. Elhayek, S. A. Ali, C. Theobalt, V. Golyanik, and D. Stricker, "HandVoxNet++: 3D hand shape and pose estimation using voxel-based neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8962–8974, Dec. 2022.
[41] W. Zhao, W. Wang, and Y. Tian, "GraFormer: Graph-oriented transformer for 3D pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20406–20415.
[42] J. Y. Chang, G. Moon, and K. M. Lee, "V2V-PoseNet: Voxel-to-voxel prediction network for accurate 3D hand and human pose estimation from a single depth map," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5079–5088.
[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.
[44] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes," in *Proc. Robot., Sci. Syst. XIV*, Jun. 2018.
[45] F. Xiong, B. Zhang, Y. Xiao, Z. Cao, T. Yu, J. T. Zhou, and J. Yuan, "A2J: Anchor-to-joint regression network for 3D articulated pose estimation from a single depth image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 793–802.
[46] A. Spurr, U. Iqbal, P. Molchanov, O. Hilliges, and J. Kautz, "Weakly supervised 3D hand pose estimation via biomechanical constraints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 211–228.

**AHMED TAWFIK ABOUKHADRA** received the B.S. degree in computer science from German University in Cairo (GUC), Egypt, in 2019, and the master's degree in artificial intelligence from Maastricht University, The Netherlands, in 2021. He is currently pursuing the Ph.D. degree. He is a Researcher with the Augmented Vision Group, DFKI, and the Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau (RPTU). His research interest includes hand-object 3D reconstruction using deep learning.

**JAMEEL MALIK** received the master's degree in electrical engineering from the School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Pakistan, and the Ph.D. degree in computer science from Technische Universit, Kaiserslautern, in 2020, for his work on depth-based 3D hand pose and shape estimation. He is an Assistant Professor at SEECS, NUST, and an Adjunct Post-Doctoral. He is currently a Postdoctoral Researcher with the Augmented Vision Group, German Research Center for Artificial Intelligence (DFKI GmbH), Kaiserslautern. His current research interests include computer vision, deep learning, and their applications.

**NADIA ROBERTINI** received the master's degree in visual computing from Saarland University, Germany, and the Ph.D. degree from the Max Planck Institute for Informatics, Saarbrücken, in 2019, for her work on full-body 3D human performance capture. She is currently a Postdoctoral Researcher with the Augmented Vision Group, German Research Center for Artificial Intelligence (DFKI GmbH), Kaiserslautern. Her current research interests include computer vision, machine learning, and human performance capture.



**AHMED ELHAYEK** received the master's degree from Saarland University, Germany, in 2010, and the Ph.D. degree from Max-Planck-Institute and Saarland University, in 2015. He was a Postdoctoral Researcher with the Augmented Vision Group, German Research Centre for Artificial Intelligence (DFKI). He joined the Faculty of Computer and Cyber Sciences, University of Prince Mugrin (UPM), as an Assistant Professor, in 2018. He founded the Artificial Intelligence Department, UPM, in 2019. He is currently the Head of the Artificial Intelligence Department.

**DIDIER STRICKER** is currently a Professor of computer science with the University of Kaiserslautern and the Scientific Director with DFKI, where he leads the Augmented Vision Research Group. His research interests are cognitive interfaces, user monitoring and on-body-sensor networks, computer vision, video/image analytics, and human–computer interaction. He received the Innovation Prize from German Society of Computer Science, in 2006, and several awards for best papers or demonstrations at different conferences. He serves as a reviewer for different European or national research organizations. He is a reviewer of different journals and conferences in the area of VR/AR and computer vision.