Graph-Based Hand-Object Meshes and Poses Reconstruction With Multi-Modal Input

MURAD ALMADANI1,2, AHMED ELHAYEK3, JAMEEL MALIK4, AND DIDIER STRICKER1
1Augmented Vision Group, German Research Center for Artificial Intelligence, 67663 Kaiserslautern, Germany
2Department of Biomedical Engineering, Khalifa University, Abu Dhabi, United Arab Emirates
3University of Prince Mugrin, Medina 42241, Saudi Arabia
4School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan
Corresponding author: Murad Almadani (murad.almadani@dfki.de)

This project is funded by the BMBF-project DECODE (01IW21001).

ABSTRACT Estimating the hand-object meshes and poses is a challenging computer vision problem with many practical applications. In this paper, we introduce a simple yet efficient hand-object reconstruction algorithm. To this end, we exploit the fact that both the poses and the meshes are graphs-based representations of the hand-object with different levels of details. This allows taking advantage of the powerful Graph Convolution networks (GCNs) to build a coarse-to-fine Graph-based hand-object reconstruction algorithm. Thus, we start by estimating a coarse graph that represents the 2D hand-object poses. Then, more details (e.g., third dimension and mesh vertices) are gradually added to the graph until it represents the dense 3D hand-object meshes. This paper also explores the problem of representing the RGBD input in different modalities (e.g., voxelized RGBD). Hence, we adopted a multi-modal representation of the input by combining 3D representation (i.e., voxelized RGBD) and 2D representation (i.e., RGB only). We include intensive experimental evaluations that measure the ability of our simple algorithm to achieve state-of-the-art accuracy on the most challenging datasets (i.e., HO-3D and FPHAB).

INDEX TERMS Hand pose estimation, hand shape estimation, hand-object interaction, graph convolution, machine learning.

I. INTRODUCTION Our hands are the main tool that we use for interaction with objects in the real world. Thus, hand-object reconstruction from a monocular image is a very important computer vision problem. Accurately estimating hand-object meshes and poses is crucial for many practical applications including virtual reality and augmented reality, human-computer interaction, fine-grained action recognition, imitation-based learning, and telepresence. Moreover, understanding hand-object interactions is essential for developing robots that perceive and act in the world. Recently, a plethora of studies focused on hand pose estimation [1]–[6]. In consequence, impressive hand pose estimation results have been achieved. On the other hand, only a limited number of studies targeted the hand-object poses estimation problem [7]. Although the hand-object mesh and pose reconstruction problem is essential for many applications, this problem has received very limited attention due to its challenging nature. This is due to the significant occlusions of both the hand and object, limited real hand-object shape datasets, varying hand shapes, and the self-similarity of hand parts. To overcome the hand-object occlusion problem, it is important to jointly estimate the hand shape along with the manipulated object. Some algorithms [1], [2] solve these two inherently coupled problems separately which leads to sub-optimal results. Several deep learning-based algorithms [8], [9] modeled the hand and object poses as a single set of 3D joints. We follow [7], and jointly model the hand and object shapes as two meshes.

As both poses and meshes are essentially two graphs-based representations of the real hand-object with different levels of details, we propose to solve the hand and object reconstruction problems simultaneously using dynamic coarse-to-fine graph convolutional neural networks (GCNs). Our network dynamically grows from a coarse 2D representation of the hand-object poses until it represents the dense 3D hand-object meshes. This is supported by the ability of the GCNs to learn effective representations of graph-structured data and its ability to encapsulate any kind of feature in the graph’s nodes. Moreover, GCNs can be used to learn the inter-joint kinematic constraints and the relationships between the vertices of the
In this section, we focus on the related works that reconstruct hand-object meshes and poses of hand-object which achieves state-of-the-art accuracy on the most challenging datasets. Due to these facts, GCNs have recently received much attention in many fields.

For all these reasons, we propose a novel dynamic GCN-based pipeline for accurate hand-object poses and meshes estimation. Going from coarse to fine, we first estimate the hand-object 2D pose (i.e. 29 2D joints) which is then converted to a 3D pose using a GCN-based network. Thereafter, the number of nodes in the graph increases gradually via multiple GCNs. The final output graph provides the deformed 3D meshes vertices of both hand and object; see Figure 1.

Although the advent of RGBD sensors brought remarkable progress to hand pose estimation, the RGBD data is still not exploited optimally for hand-object reconstruction. In this study, we explore the problem of representing the RGBD input in different modalities. One of these modalities is the voxelized RGBD representation where each voxel contains the value of a truncated signed distance function (TSDF) [10] and the color value of the corresponding RGB pixel. This 3D representation allows our 3D encoder to extract better features. Our final pipeline adopts a multi-modal representation of the input by combining 3D representation (i.e. voxelized RGBD) and 2D representation (i.e. RGB only).

In this paper, we propose a novel end-to-end framework for recovering full hand-object meshes and poses based on coarse-to-fine GCNs and multi-modal representation of the RGBD input image. The main contributions of our work are summarized as follow:

- A coarse-to-fine Graph-based hand-object reconstruction network.
- A multi-modal representation of the input RGBD image which combines voxelized RGBD features with 2D features.
- A novel end-to-end framework for recovering full 3D meshes and poses of hand-object which achieves state-of-the-art accuracy on the most challenging datasets.

II. RELATED WORK

In this section, we focus on the related works that reconstruct both hand and object from monocular input. We refer the reader to [11], [12] for a detailed overview of works focusing on the reconstruction of hands and objects in isolation.

A. HAND-OBJECT INTERACTION

The problem of vision-based hand-object reconstruction from a single image has been addressed from several perspectives. One stream of works in this direction focuses on a hand alone, e.g. hand pose inferring and RGB or RGBD data configuration to either control the hand [13]–[15] or infer manipulation behavior [16]–[18]. Another line of study incorporates the concept that an object’s structure determines the hand pose and investigates the hand along with the object [19]–[21]. Some researchers also seek to predict the forces and interactions between the hand and the manipulated object [22]. Scaling this approach to many objects, nevertheless, would require large RGB or RGBD datasets annotated with poses of hands and objects that are costly to gather. Due to this reason, existing datasets either contain a small number of instances annotated with a 6D pose [23] or a large number of instances with coarse annotations such as grasp types, or bounding boxes [24], [25], making them insufficient for an in-depth understanding of the interactions between hand and object. In this work, we propose an approach for estimating both the hand and the object pose and shape from a single RGBD image, which could theoretically enable us to automatically annotate and use these datasets for further study of physical hand-item interactions.

B. HAND-OBJECT POSE AND SHAPE ESTIMATION

Most approaches addressed the problem of hand-object reconstruction by estimating the hand pose and the object pose separately. For example, Brachman et al. [26] used a multi-stage approach to recover 6D object pose from a single RGB image using regression forests. Both [27], [28] are based on 2D/3D Convolutional Neural Networks (CNNs). BB8 [28] used 2D CNNs to segment the object, and predicted it’s 3D bounding box. The object’s 6D pose is then estimated via PnP [29]. [27] predicted the objects pose along with its 2D bounding boxes. The approaches mentioned above require a detailed 3D object model with texture as input. They also require a further pose refinement step to improve their precision. Tekin et al. [1] resolved these limitations by implementing a single-shot architecture that forecasts 2D projections of the object’s 3D bounding box in a single forward pass.

All these approaches do not address the problem of object prediction in hand-object interaction scenarios where objects could be largely obscured. Recently, the 3D hand pose and shape estimation problem got more attention [2], [9], [30]–[33]. However, this problem is challenging [34], [35] due to self-occlusions and the limited amount of datasets available [32], [33]. Mueller et al. [32] used 2D CNNs on synthetic data and integrated it with a generative hand model to detect hands dealing with objects from RGBD images. Their heuristic model was enhanced to work with RGB videos [36]. Iqbal et al. [37] estimated 3D hand pose from a single RGB image (both from egocentric and third-person...
viewpoints) via CNN. On the other hand, several works [11], [38]–[41] utilized only depth maps instead of RGB images for estimating hand poses. However, many of the existing approaches focus on predicting the hand pose and aim to be stable in the presence of objects, but do not address the problem of simultaneous estimation of both hand and object. The approaches in [30], [32] indiscriminately inferred the 3D hand pose from the depth image by finding that different object shapes cause different hand grasps. Model-based techniques to recover hand and object parameters were also proposed [42]. Most strategies, nevertheless, concentrated on third-view scenarios [14], [43], [44].

C. GRAPH CONVOLUTIONAL NETWORKS

The state-of-the-art Graph-convolution networks enable the development of high-level interpretations of connections among graph-based data nodes. For 2D and 3D human pose estimation, Zhao et al. [45] introduced a semantic graph convolution network to capture local as well as global relations between human body joints. By encoding domain information of the human body and hand joints, Cai et al. [46] transformed 2D human joints to 3D using a graph convolution network that can develop multi-scale models. For skeleton-based action recognition, Yan et al. [47] leveraged a graph convolution network to acquire a spatial-temporal analysis of human body joints. Kolotouros et al. [48] described that GCNs may be exploited to acquire 3D human shape and pose from a single RGB image. Whereas, Ge et al. [49] used the graph convolution to estimate the full 3D hand mesh from a monocular RGB image. Similarly, authors in [7] proposed their Adaptive Graph U-Net to estimate the key-points of the hand-object interaction from a single RGB image. They leveraged their algorithm from the Graph U-net which was firstly presented in [50]. They found that the pooling and unpooling layers in the original Graph U-Net didn’t work well enough on graphs with low edge numbers, such as skeletons or object meshes.

In this paper, we utilize the new trainable Graph U-Net architecture used in [7], and we propose a novel coarse-to-fine GCNs based end-to-end network to simultaneously estimate the hand-object poses and shapes using multi-modal input. Our solution inherently preserves the structure of the estimated hand and object meshes.

III. METHOD OVERVIEW

Given an input RGBD image, our goal is to estimate 3D hand pose \( \mathcal{P}_{3D} \), 3D object pose \( \mathcal{P}_{3D}^o \), hand mesh (i.e. shape) \( \mathcal{V}_{out}^{h} \), and object mesh \( \mathcal{V}_{out}^{o} \). For simplicity, we use \( \mathcal{P}_{3D} \) and \( \mathcal{V}_{out} \) to represent the hand-object poses and meshes, respectively. \( \mathcal{P}_{3D} \) consists of \( N = 29 \) joint locations \( \mathcal{J} \in \mathbb{R}^{3 \times N} \), while \( \mathcal{V}_{out} \) consists of \( K = 1778 \) 3D vertex locations \( \mathcal{V} \in \mathbb{R}^{3 \times K} \). Fig. 2 shows an overview of our pipeline that simultaneously reconstructs hand-object in interaction.

The RGBD input is transformed into a voxelized grid representation \( \mathcal{V}_{RGBD} \) in two steps. First, the depth input is transformed into a voxelized grid (i.e., \( \mathcal{V}_{D} \)) of size \( 88 \times 88 \times 88 \), by using intrinsic camera parameters, a fixed cube size, and the truncated signed distance function (TSDF). Then, \( \mathcal{V}_{RGBD} \) is generated by adding the color information to the corresponding voxels. The 3D encoder extracts the features \( f_2 (8000D) \) from \( \mathcal{V}_{RGBD} \) by applying 3D convolutions on \( \mathcal{V}_{RGBD} \). At the same time, the 2D encoder extracts feature vector \( f_1 (2000D) \) from the RGB image. A multi-modal feature vector \( \mathcal{F} (1000\text{D}) \) is generated by concatenating \( f_1 \) and \( f_2 \). \( \mathcal{F} \) is provided as an input to the 2D pose estimator in order to generate the hand-object 2D poses. Once this 2D pose is estimated, graph convolution networks are used in all the following steps as this pose represents the first (i.e. coarse) graph.

Our course-to-fine Graph-based network needs enough information to be able to produce finer details from coarse inputs. To this end, the node of any graph is generated by concatenating the multi-modal feature vector \( \mathcal{F} \) with either the coordinates of the 2D joint, 3D joints, or mesh vertices. Give such rich information in each node of the graph, the GCNs can generate finer hand-object representations with high accuracy. In Stage 1, a Graph-based 2D pose refinement network is applied to improve the accuracy of the first 2D pose. Then, another Graph-based 3D pose estimator (i.e. Adaptive Graph U-Net [7]) generates the 3D poses \( \mathcal{P}_{3D} \) of hand-object. Finally, in Stage 2, a novel graph-based Hand-object shape generator is used to jointly produce full 3D meshes of hand and object \( \mathcal{V}_{out} \) given the 3D poses as input.

Our loss function consists of four different parts. Firstly, the loss for the initial 2D coordinates predicted by the 2D pose estimator is calculated (\( \mathcal{L}_{2D_{\text{out}}} \)). Thereafter add the refined 2D pose (\( \mathcal{L}_{2D} \)), 3D pose (\( \mathcal{L}_{3D} \)), and 3D mesh coordinates \( \mathcal{L}_{out} \) losses, respectively, as follow:

\[
\mathcal{L} = \lambda_{2D} \mathcal{L}_{2D_{\text{out}}} + \lambda_{2D} \mathcal{L}_{2D} + \lambda_{3D} \mathcal{L}_{3D} + \lambda_{V_{out}} \mathcal{L}_{V_{out}}
\]

where the losses are calculated using Mean Squared Error function. The weights \( \lambda_{2D} = 0.1, \lambda_{3D} = 1, \) and \( \lambda_{V_{out}}=1 \) were found experimentally and are kept constant in all experiments.

IV. GRAPH-BASED HAND-OBJECT RECONSTRUCTION

In this section, we highlight the function and effectiveness of the main components of our simple yet effective algorithm. To this end, we first propose the multi-modal representation of the RGBD input which allows improving the reconstruction accuracy. Thereafter, we discuss the structure of our dynamic course-to-fine GCNs which exploit the graph nature of the poses and the meshes to generate accurate and robust hand-object reconstruction.

A. MULTI-MODAL INPUT REPRESENTATION

In this subsection, we explore the problem of representing the RGBD input in different modalities. One of the possible modalities is the TSDF-based voxelized representation which transforms the RGBD input into voxelized grid representation. The advantage of this representation is two-fold. First, the depth map is inherently 2.5D data which can be better...
FIGURE 2. Overview of our pipeline that simultaneously reconstructs hand-object in interaction. The input to our algorithm is an RGBD image that is transformed into a voxelized grid representation \( V_{RGBD} \). The 3D encoder extracts 3D features \( f_2 \) from \( V_{RGBD} \). Also, the 2D encoder extracts the feature vector \( f_1 \) from an RGB image. A multi-modal feature vector \( F \) is generated by concatenating \( f_1 \) and \( f_2 \). \( F \) is provided as an input to the following stages.

In Stage 1, a Graph-based pose estimator jointly generates the 3D pose of hand and object \( \mathcal{P}_{3D} \). Finally, Stage 2 predicts meshes of both hand and object \( V_{out} \) using a coarse-to-fine Graph-based shape estimator network.

represented in a 3D voxelized grid using binary quantization (occupancy grid) [11], [38] or TSDF [51]. The TSDF-based representation is more effective than the occupancy grid because TSDF allows to better encode the depth information by recognizing the voxels before and behind the observed surface [51]. The TSDF-based representation of the depth map can be further enriched by adding color information to the voxels [10]. Second, 3D volumetric representation of depth data allows using 3D convolutions which better extract the features as compared to 2D convolutions [38].

The input depth map is first transformed into a 3D point cloud using camera intrinsics. These points are then discretized in the range \([1, 88]\). A voxelized grid of size \( 88 \times 88 \times 88 \) is created by evaluating \( V(n) \) for every voxel \( n \). In the TSDF representation, \( V(n) \) can be obtained as:

\[
V(n) = \min(\max(D(n_c)/\mu, -1), +1),
\]

where \( D(n_c) \) is the signed distance from the voxel center \( n_c \) to the closest surface point in the discretized 3D point cloud. Here, we use the euclidean distance. The sign of \( D(n_c) \) is positive if the depth value of \( n_c \) is less than the depth value of the closest surface point else, the sign is negative. We select \( \mu = 3 \) as the truncated distance value.

We employ the projective TSDF [52] which is practically more feasible because the closest surface point is found on the line of sight in the camera frame.

1) ADDING COLOR INFORMATION TO THE VOXELIZED GRID
Given the calibration between RGB channel and depth channel (captured by two separate sensors), we get the RGB color value corresponding to each pixel in the depth image. Once this correspondence is known, the color value of any RGB pixel is concatenated with the TSDF value of all voxels influenced by the corresponding depth pixel.

2) OTHER MODALITIES
There are several possible representations of the RGBD input. One of them is the traditional representation where the RGBD
data is represented as a 2D image with 4 channels. Another possibility is to use 3D representation (e.g., TSDF or occupancy grid) for the depth channel and 2D representation of the RGB channels.

For the best accuracy, we adopted a multi-modal representation of the input by combining the VRGBD 3D representation (i.e., voxelized RGBD) and the color 2D representation (i.e., RGB image). As mentioned above, these two representations are generated using the 2D encoder and the 3D encoder (i.e., encoder of the V2V-PoseNet [38]) to generate the multi-modal feature vector $\mathcal{F}$. The 2D encoder is a ResNet10 which is a lightweight version of the original residual neural network [53].

### B. 3D HAND AND OBJECT POSE ESTIMATION

Our 3D pose estimation network (i.e., Stage 1 of Fig. 2) is inspired by the architecture of [7] which is based on RGB input. In contrast, our 3D pose estimation network is adapted to work with multi-modal input. Moreover, the rich multi-modal feature vector $\mathcal{F}$ is embedded in the nodes of the graphs at each stage of the network to improve hand-object poses reconstruction accuracy; see Sec. V-C.

In the first step of Stage 1, an initial 2D hand-object poses are estimated by the 2D pose estimator. This estimator is a simple fully connected layer that converts the multi-modal feature vector $\mathcal{F}$ to the 2D hand pose (i.e., 25 joint locations) and the 2D pose of the object (4 corners of the object’s bounding box). Then, the 2D pose refinement network (i.e., 3 graph convolution layers) utilizes the power of neighbors features to improve the estimation of the 2D pose. To this end, $\mathcal{F}$ is concatenated with the coordinates of the hand-object 2D poses to form the input of the refinement network. The encoder-decoder based Adaptive Graph U-Net was used in [7] for lifting the refined 2D poses to 3D. We use the same number of graph U-stages as in [7]. However, [7] do not provide any clue of the 3rd dimension to the graph U-Net to support this conversion. In contrast, we concatenate our multi-modal feature vector $\mathcal{F}$ with the refined 2D pose to improve the accuracy; see Section V-B.

### C. 3D HAND AND OBJECT SHAPE ESTIMATION

In this subsection, we propose our novel Hand-object Shape Generator network that estimates 3D meshes given 3D pose. This network preserves the structures of the meshes as it is based on graph convolutions. Other works such as [11], [49], [54], [55] have estimated 3D hand meshes from 2D (or 3D) hand pose. However, they do not explicitly encode the structure of hand meshes because they are based on 2D (or 3D) convolutions. The input of Stage 2 (i.e., Hand-object Shape Generator network) is a graph generated by concatenating our multi-modal feature vector $\mathcal{F}$ with the coordinates of 3D poses.

As shown in figure 2, the number of nodes in the graph increases gradually from 29 to 112, 445, and 1778 vertices via three sets of unpoolings-GCN layers. Each of these sets starts by one unpooling layer followed by two Graph CNN layers.

We use the Chebyshev Spectral Graph CNN layers that have been firstly proposed in [56]. The intermediate graphs are defined using the quadric edge collapse decimation algorithm [57], and fixed during training and testing.

As part of our training, we adopted four different losses incorporated in the final mesh loss ($\mathcal{L}_{\text{out}}$), as follows:

$$
\mathcal{L}_{\text{out}} = \lambda_v \mathcal{L}_v + \lambda_n \mathcal{L}_n + \lambda_e \mathcal{L}_e + \lambda_l \mathcal{L}_l,
$$

where $\lambda_v$, $\lambda_n$, $\lambda_e$ and $\lambda_l$ are regularization parameters, and $\mathcal{L}_v$, $\mathcal{L}_n$, $\mathcal{L}_e$, and $\mathcal{L}_l$ are the vertex loss, normal loss, edge loss, and the Laplacian loss, respectively; see [58] for more details.

### V. EXPERIMENTS

In this section, we first introduce the evaluation datasets followed by a comparison with the state-of-the-art methods. Then, we conduct comprehensive ablation studies on various input modalities.

### A. DATASETS

We used two datasets to evaluate our hand-object poses and shapes estimation method, namely the first-person hand action benchmark (FPHAB) [23] and the HO-3D [59]. FPHAB dataset offers 11,019 frames for training, and 10,482 for evaluation. We choose our training and evaluation set using the action split protocol that has been used in [23]. Each frame’s annotation is a 6D vector that provides 3D translation and rotation for each of the objects. To fit this annotation to our graph model, we translate and rotate the 3D object mesh to the annotation’s pose for each object in each frame, and then compute a tight oriented bounding box by applying PCA on vertices. In our graph, we use the eight 3D coordinates of the object box corners as nodes.

HO-3D dataset contains video sequences of hands and handled objects captured from a third-person viewpoint. Along with ten subjects and ten objects taken from the YCB [60] dataset, the provided RGBD frames were annotated for hand-object poses and meshes. The training set consists of 66,034 frames while the evaluation set consists of 11,524 frames. Since a web challenge is created for HO-3D dataset, the ground truth annotations of HO-3D are only available for its training set.

**Evaluation Metrics:** We use two evaluation metrics: (i) the average 3D keypoints location error over all test frames (3D P Err.); (ii) mean vertex location error over all test frames (3D V Err.).

### B. EVALUATION OF HAND RECONSTRUCTION

Firstly, in order to evaluate our framework for hand pose and shape estimation, we contribute to the corresponding HO-3D challenge [59]. The challenge fairly assesses frameworks’ ability to estimate hand pose and shape with the

1. https://competitions.codalab.org/competitions/22485?#results
presence of four different manipulated objects. Table 1 shows the quantitative results of the two most recent published frameworks that contributed to the challenge along with the results of our framework. It shows that our method is the only method that utilizes both RGB image and depth maps. This utilization reflects on enhancing the produced average error in the constructed hand pose and shape. The table shows that we outperform the Spurr et al. [61] and V2V posenet [38] frameworks by 7 mm and 13.1 error in hand pose reconstruction, respectively. Whereas, we improve the hand shape reconstruction error by 16.5 mm as we compare with the V2V posenet [38]. Except for the results related to the V2V posenet, which we generated using their publicly available code, all other quantitative results can be found on the HO-3D leaderboard challenge webpage. When comparing our method to the other mentioned pipelines, we are the only one that uses GCNs. This emphasizes how important it is to take advantage of the fact that both pose and mesh are graph-based representations, which outperform the well-known V2V [38] framework, which achieved promising results on the HANDS2017 [62] hand reconstruction challenge.

On the other hand, since HO-3D provides ground truth annotations only for its training set, and in order to evaluate our framework qualitatively, we select approximately 80% of the original training data (i.e., 52,827 frames) as our new train set and the remaining data (i.e., 13,207 frames) as our test set. For fair qualitative evaluation, we split the new test and training sets such that we got the same testing errors as we got on the challenge webpage. Figure 3 shows six different hand pose and shape reconstructed samples of the new split testing set. In this figure, we present the estimated hand shape in green color aligned with the ground truth in gray color, whereas, we show the corresponding pose estimation and ground truth in red and blue colors, respectively. This figure shows the robustness of our framework to well reconstruct hand pose and shape concerning its corresponding ground truth.

### C. EVALUATION OF HAND-OBJECT RECONSTRUCTION

In the previous subsection, we show the evaluation results of only hand reconstruction since the HO-3D challenge doesn’t provide object reconstruction results. However, in this section, we show the results of complete hand and object reconstruction using our framework. Figure 4 shows four different reconstruction samples of our split HO-3D test set. Since our method calculates the complete 2D pose, 3D pose, and 3D mesh throughout the hand-object reconstruction, we show in this figure the complete reconstruction steps of our framework. The ground truth annotations are shown in the upper row of each sample, while the estimates are shown in each second row. We show the 2D pose projection on its corresponding RGB image along with two different viewpoints of both 3D pose and shape for better visualization comparison. The figure shows the robustness of our framework to well calculate all pipeline stages throughout the full reconstruction. Also, we provide more visualizations of our hand-object reconstruction for the HO-3D dataset in figure 5.
This figure also shows two different points of view for the final hand-object reconstruction with a variety of interaction scenarios, which bring to light the robustness of our reconstruction method with minor errors.

We also test our framework on the FPHAB [23] dataset, which is the second type of hand-object interaction dataset. Unlike the HO-3D dataset, FPHAB assesses the framework's ability to reconstruct hand-object pose interactions with an egocentric point of view. Table 2 displays the quantitative hand-object reconstruction results using three of the most recent published frameworks, as well as our own. We only provide results for 3D hand and object pose estimation separately because FPHAB does not provide ground truth annotation of hand meshes. Our algorithm outperforms Hasson et al. [63], Tekin et al. [8], and the baseline [23] by 7.7 mm, 5.5 mm, and 4 mm in hand pose reconstruction, respectively. Whereas, in comparison to Tekin et al. [8] and Hasson et al. [63], we improve object pose estimation by 13.4 mm and 10.8 mm, respectively. The table also shows the input modalities used in the mentioned frameworks, indicating that only ours and baseline use both RGB and Depth maps in the reconstruction. However, unlike the baseline framework, we use a 3D representation of the depth map, which results in a significant improvement in hand pose reconstruction. In addition, as shown in Figure 6, we provide qualitative evaluation results on FPHAB hand-object pose reconstruction. The projected ground truth annotation for various hand-object interaction scenarios are shown on top of the corresponding estimations in this figure. In the case of an egocentric point of view, it shows that our method is stable in both simple and complex interaction scenarios.

**Runtime:** Table 3 shows the runtime of each component of our pipeline. We evaluate the runtime of our method on the Nvidia RTX 2070 GPU. It can be seen that stage 2 and the input generation consume the majority of the running time (i.e. 128 ms). The rest of the time was spent in stage 1 and extracting the input features (i.e. 68 ms).

**D. INPUT MODALITIES ABLEATION STUDY**

In this section, we demonstrated experimental studies of our entire framework using multi-input modalities (both RGB and voxelized RGBD). Thus, we present comprehensive ablation studies on various input modalities using the HO-3D evaluation dataset. To this end, we compare our entire

---

**TABLE 2. Quantitative results on FHPAB [23] dataset.** We show a side-by-side comparison with the three most recent methods for hand and object pose reconstruction. Our method considers the first study to utilize both RGB and Depth maps and results with the smallest reconstruction average error.

| Methods            | 3D Hand Err (mm) | 3D Object Err (mm) |
|---------------------|------------------|-------------------|
| Hasson et al. [63]  | 18.0             | 22.3              |
| Tekin et al. [8]    | 15.8             | 24.9              |
| Baseline [23] (RGBD)| 14.3             | –                 |
| Ours (RGBD)         | 10.3             | 11.5              |

**TABLE 3. Runtime.** We show the detailed runtime of each component of our pipeline.

| Component                     | Runtime (sec) |
|-------------------------------|---------------|
| Multi-modal Input generation  | 0.061         |
| Feature extraction            | 0.015         |
| Stage 1                       | 0.053         |
| Stage 2                       | 0.067         |
| Total                         | 0.197         |
pipeline to the three pipelines shown in figure 7. The top pipeline is based on the voxelized RGBD modality (i.e. without RGB modality which allows extracting features from RGB using a 2D encoder). The middle row of figure 7 illustrates the pipeline which totally neglects the RGB data. The pipeline of the bottom row allows investigating the effect of using a raw depth map as an input rather than the voxelized TSDF. Table 4, summarize the hand reconstruction errors for four pipelines (i.e. the three pipelines of figure 7 and the RGB only pipeline which is very similar to the pipeline of the Raw depth map). This error was estimated using the HO-3D challenging dataset. When compared to the 2D raw depth map, the use of the voxelized TSDF input has a significant impact on the reconstructed errors, reducing them by approximately 2.5 mm and 7.2 mm for the hand pose and shape, respectively. This lighting on the novelty of our method which employs 3D convolutional layers to extract features from the depth information. Similarly, the table shows that using the voxelized RGBD instead of the voxelized TSDF improves pose and shape by 0.1 mm and 0.2, respectively. The RGB-only pipeline yields the lowest accuracy as it lacks 3D information. Finally, it demonstrates that using the full multi-modal input outperforms other methods, with a final error of 11.4 mm for both pose and shape hand reconstructions. Please note that it is essential to concatenate the multi-modal feature vector with each node of the GCN in all stages of our pipeline. Otherwise, the 3D hand reconstitution fails.

### Table 4. Ablation study. Quantitative comparison between four different input modalities on HO-3D [59].

| Input Model         | 3D Pos. Err. (mm) | 3D V. Err. (mm) |
|---------------------|-------------------|-----------------|
| Raw depth Map       | 14.0              | 18.8            |
| Voxelized TSDF      | 11.5              | 11.6            |
| Voxelized RGBD      | 11.4              | 11.5            |
| RGB only            | 16.2              | 16.2            |
| Voxelized RGBD + RGB (Ours) | **11.4** | **11.4** |

### VI. CONCLUSION AND FUTURE WORK

We introduce a novel Graph-based deep network for accurate reconstruction of hand-object meshes and poses using a single RGBD input. Our dynamic GCNs grow from a...
coarse 2D representation of the hand-object poses until it represents the dense 3D hand-object meshes. This is supported by the ability of the GCNs to encapsulate rich features in the graph’s nodes. This joint Graph-based coarse-to-fine strategy produces more accurate hand and objects meshes. The experimental evaluation shows that the TSDF-based voxelized representation of RGBD input allows obtaining better features. These features when combined with the features extracted from a single RGB image further enhance the accuracy of the reconstructions. We achieve state-of-the-art results for hand-object meshes and poses reconstruction, which is confirmed on recent challenging benchmarks. In future work, a discriminator network can be used to determine whether the generated meshes correspond to real hand-object meshes, which can lead to improved accuracy of the reconstructions. Another possible extension of our algorithm is to add a temporal constraint on the hand-object reconstruction.

ACKNOWLEDGMENT

This project is funded by the BMBF-project DECODE (01IW21001).

REFERENCES

[1] B. Tekin, S. N. Sinha, and P. Fua, “Real-time seamless single shot 3D object pose prediction,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 292–301.

[2] S. Yuan, G. Garcia-Hernando, B. Stenger, G. Moon, J. Y. Chang, K. M. Lee, P. Molchanov, Kautz, J. S., H. Sonari, L. Ge, and J. Yuan, “Depth-based 3D hand pose estimation: From current achievements to future goals,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 2636–2645.

[3] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit, “HONote: A method for 3D annotation of hand and object poses,” 2019, arXiv:1907.01481. [Online]. Available: https://arxiv.org/abs/1907.01481

[4] D. Goudie and A. Galata, “3D hand-object pose estimation from depth with convolutional neural networks,” in Proc. 12th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG), May 2017, pp. 406–413.

[5] J. Malik, A. Elhayek, F. Nunnari, and D. Stricker, “Simple and effective deep hand shape and pose regression from a single depth image,” Comput. Graph., vol. 85, pp. 85–91, Dec. 2019.

[6] M. R. Al Koutanyi, V. Rybalkin, J. Malik, A. Elhayek, C. Weiss, G. Reis, N. Wehn, and D. Stricker, “Real-time energy efficient hand pose estimation: A case study,” Sensors, vol. 20, no. 10, p. 2828, May 2020.

[7] 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., Jun. 2020, pp. 6608–6617.

[8] 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. 4511–4520.

[9] Y. Hasson, G. Varol, I. Kalevatykh, M. J. Black, I. Laptev, and T. Feix, I. M. Bullock, and A. M. Dollar, “Analysis of human grasping behavior: Object characteristics and grasp type,” IEEE Trans. Haptics, vol. 7, no. 3, pp. 311–323, Jul. 2014.

[10] M. Cai, K. M. Kitani, and Y. Sato, “Understanding hand-object manipulation with grasp types and object attributes,” in Robotics: Science and Systems. 2016.

[11] M. Cai, K. M. Kitani, and Y. Sato, “An ego-vision system for hand grasp analysis,” IEEE Trans. Human-Mach. Syst., vol. 47, no. 4, pp. 524–535, Aug. 2017.

[12] G. Rogez, J. S. Supanic, and D. Ramanan, “Understanding everyday hands in acts,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 3889–3897.

[13] 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.

[14] D. Damen et al., “Scaling egocentric vision: The EPIC-KITCHENS dataset,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 720–736.

[15] Y. Li, Z. Ye, and J. M. Rehg, “Delving into egocentric actions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 287–295.

[16] E. Brachmann, F. Michel, A. Krull, M. Y. Yang, S. Gumhold, and C. Rother, “Uncertainty-driven 6D pose estimation of objects and scenes from a single RGB image,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 3364–3372.

[17] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, “SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 1521–1529.

[18] M. Rad and V. Lepetit, “BB8: A scalable, accurate, robust to partial occlusion method for predicting the 3D poses of challenging objects without using depth,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 3828–3836.

[19] V. Lepetit, F. Moreno-Noguer, and P. Fua, “EPnP: An accurate o (n) solution to the PNP problem,” Int. J. Comput. Vis., vol. 81, no. 2, p. 155, 2009.

[20] C. Choi, S. H. Yoon, C.-N. Chen, and K. Ramani, “Robust hand pose estimation during the interaction with an unknown object,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Nov. 2017, pp. 3123–3132.

[21] 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. 1154–1163.

[22] G. Rogez, J. S. Supanic, and D. Ramanan, “First-person hand pose recognition using egocentric workspaces,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 4325–4333.

[23] Q. Ye and T.-K. Kim, “Occlusion-aware hand pose estimation using hierarchical mixture density network,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 801–817.

[24] A. Spurr, J. Song, S. Park, and O. Hilliges, “Cross-modal deep variational pose estimation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 11807–11816.

[25] C. Zimmermann and T. Brox, “Learning to estimate 3D hand pose from single RGB images,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 4903–4911.

[26] 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.
U. Iqbal, P. Molchanov, T. B. J. Gall, and J. Kautz, “Hand pose estimation via latent 2.5D heatmap regression,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 118–134.

J. Y. Chang, G. Moon, and K. M. Lee, “V2 V-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.

J. Malik, A. Elhayek, and D. Stricker, “Simultaneous hand pose and skeleton bone-lengths estimation from a single depth image,” in Proc. Int. Conf. 3D Vis. (3DV), Oct. 2017, pp. 557–565.

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.

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, 2018.

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.

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), Sep. 2018, pp. 484–500.

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, 2016.

L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, “Semantic graph convolutional networks for 3D human pose regression,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 3425–3435.

Y. Cai, L. Ge, J. Liu, J. Cai, T.-J. Cham, J. Yuan, and N. M. Thalmann, “Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 2277–2281.

S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” 2018, arXiv:1801.07455. [Online]. Available: https://arxiv.org/abs/1801.07455

N. Kolotouros, G. Pavlakos, and K.Daniilidis, “Convolutional mesh regression for single-image human shape reconstruction,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 4501–4510.

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. 10833–10842.

H. Gao and S. Ji, “Graph U-Nets,” 2019, arXiv:1905.05178. [Online]. Available: https://arxiv.org/abs/1905.05178

L. Ge, H. Liang, J. Yuan, and D. Thalmann, “Real-time 3D hand pose estimation with 3D convolutional neural networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 4, pp. 956–970, Apr. 2019.

R. A. Newcombe, A. Fitzgibbon, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, and S. Hodges, “KinectFusion: Real-time dense surface mapping and tracking,” in Proc. 10th IEEE Int. Symp. Mixed Augmented Reality, Oct. 2011, pp. 127–136.

K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016.

S. Baek, K. I. Kim, and T.-K. Kim, “Pushing the envelope for RGB-based dense 3D hand pose estimation via neural rendering,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 1067–1076.

J. Malik, A. Elhayek, and D. Stricker, “WHISP-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.

M. Defferrard, X. Bresson, and P. Vandergheynst, “Convolutional neural networks on graphs with fast localized spectral filtering,” 2016, arXiv:1606.09375. [Online]. Available: https://arxiv.org/abs/1606.09375

M. Garland, “Quadric-based polygonal surface simplification,” Carnegie-Mellon Univ. Pittsburgh PA School Comput. Sci., Pittsburgh, PA, USA, Tech. Rep., 1999.

N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang, “Pixel2Mesh: Generating 3D mesh models from single RGB images,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 52–67.

S. Hampali, M. Rad, M. Oberweger, and V. Lepetit, “HOutote: A method for 3D annotation of hand and object poses,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 3196–3206.

B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar, “Benchmarking in manipulation research: The YCB object and model set and benchmarking protocols,” 2015, arXiv:1502.03143. [Online]. Available: http://arxiv.org/abs/1502.03143

A. Spurr, U. Iqbal, P. Molchanov, O. Hilliges, and J. Kautz, “Weakly supervised 3D hand pose estimation via biomechanical constraints,” 2020, arXiv:2003.09282. [Online]. Available: https://arxiv.org/abs/2003.09282

S. Yuan, Q. Ye, G. Garcia-Hernando, and T.-K. Kim, “The 2017 hands in the million challenge on 3D hand pose estimation,” 2017, arXiv:1707.02237. [Online]. Available: https://arxiv.org/abs/1707.02237

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. 571–580.

MURAD ALMADANI received the B.Sc. degree in biomedical engineering from Hashemite University, Jordan, in 2013, and the master’s degree in electrical engineering from King Fahd University of Petroleum and Minerals, Saudi Arabia, in 2019. He is currently pursuing the Ph.D. degree in biomedical engineering with Khalifa University, United Arab Emirates. He is also a Research Assistant with the German Research Center for Artificial Intelligence (DFKI), Germany. His current research interests include signal and image processing, computer vision, and artificial intelligence.

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 worked as 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.

JAMEEL MALIK 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 worked as a Postdoctoral Researcher with the Augmented Vision Group, German Research Centre for Artificial Intelligence (DFKI). 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.

DIDIER STRICKER is currently a Professor of computer science with Technische Universit at Kaiserslautern and the Scientific Director of the German Research Center for Artificial Intelligence (DFKI GmbH). Kaiserslautern, where he leads the Research Department ‘Augmented Vision.’ His research interests include cognitive interfaces, user monitoring, and on-body-sensor networks, computer vision, video/image analytics, and human–computer interaction. He received the Innovation Prize of the German Society of Computer Science in 2006. He got 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.