Dynamic Hyperbolic Attention Network for Fine Hand-object Reconstruction

Zhiying Leng[1,2], Shun-Cheng Wu[2], Mahdi Saleh[2], Antonio Montanaro[3], Hao Yu[2], Yin Wang[1],
Nassir Navab[2], Xiaohui Liang[1,4*], Federico Tombari[2]
[1] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, China
[2] Computer Aided Medical Procedures, Technical University of Munich, Germany
[3] Politecnico di Torino, Italy
[4] Zhongguancun Laboratory, Beijing, China
{zhiyingleng,liang_xiaohui}@buaa.edu.cn, {shuncheng.wu,m.saleh}@tum.de, tombari@in.tum.de

Abstract

Reconstructing both objects and hands in 3D from a single RGB image is complex. Existing methods rely on manually defined hand-object constraints in Euclidean space, leading to suboptimal feature learning. Compared with Euclidean space, hyperbolic space better preserves the geometric properties of meshes thanks to its exponentially-growing space distance, which amplifies the differences between the features based on similarity. In this work, we propose the first precise hand-object reconstruction method in hyperbolic space, namely ***D****ynamic* ***H****yperbolic* ***A****ttention* ***Net****work (****DHANet****), which leverages intrinsic properties of hyperbolic space to learn representative features. Our method that projects mesh and image features into a unified hyperbolic space includes two modules,* i.e. *dynamic hyperbolic graph convolution and image-attention hyperbolic graph convolution. With these two modules, our method learns mesh features with rich geometry-image multi-modal information and models better hand-object interaction. Our method provides a promising alternative for fine hand-object reconstruction in hyperbolic space. Extensive experiments on three public datasets demonstrate that our method outperforms most state-of-the-art methods.*

1. Introduction

3D hand-object reconstruction from monocular RGB images is a fundamental task in computer vision. Given a single RGB image of a hand interacting with an object, it aims at predicting a 3D mesh of both the hand and the object under the correct pose and precisely modeling the hand-object interaction. Although the 3D posed reconstruction has a wide application in human-machine interaction, robotic grasping/learning, and augmented reality, the challenges of this task still remain in two aspects: 1) Reconstructing meshes with the pose and scale consistent with the input; 2) Fulfilling the physiological rules on hands and physical characteristics of hand-object interaction.

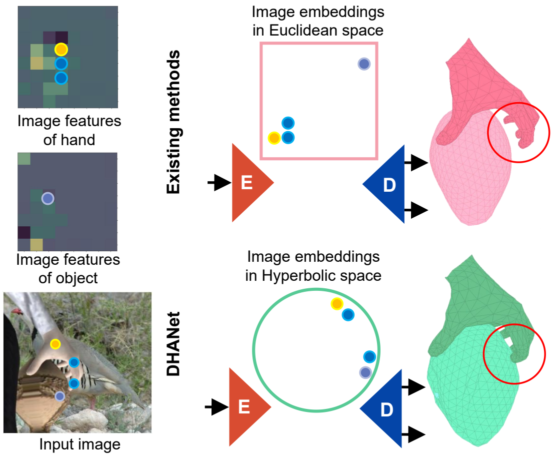

Figure 1. Colored dots indicate the features of the hand and the object. Existing methods learn image features in Euclidean space but struggle to model the contact region, resulting in separated colored dots. Our DHANet learns mesh and image features in hyperbolic space, better modeling the hand-object interaction that ensures the distribution of colored dots preserves the hand-object contact relationship. And our method results in a more accurate reconstruction, highlighted by the red circles.

Existing methods deal with hand-object images or meshes in Euclidean space [17, 28, 21, 15, 9, 51, 42, 46, 5, 14], which learn image features and regress model parameters of hand and object from Euclidean embeddings. To accurately reconstruct meshes of hands and objects, especially around the area of mutual occlusion, existing methods [28, 21, 15, 42, 9, 51] optimize the reconstruction by taking the physical interaction between the hand and the object as a cue. These methods can be broadly divided into

[*] corresponding author

two categories: learning-based methods and optimization-based methods. Learning-based methods employ attention mechanism [28, 21, 15], and other advanced models [42, 9, 51] to model hand-object interactions. Optimization-based methods integrate physical constraints, like Spring-mass System [46] and 3D contact priors [5, 14] with contact loss functions, to constraint the optimization process. Existing methods almost directly regress the model parameters of hand-object meshes from image features and manually define interaction constraints without exploiting the geometrical information. In this work, we seek for learning geometry-image multi-modal features in hyperbolic space to reconstruct accurate meshes.

As mentioned in recent research on Representation Learning in hyperbolic space [23, 2, 30, 34, 47], the effectiveness of Euclidean space for graph-related learning tasks is still bounded, failed to provide powerful geometrical representations. Compared to Euclidean space, hyperbolic space exhibits the potential to learn representative features. Due to the exponential growth property of hyperbolic space, it is innately suitable to embed tree-like or hierarchical structures with low distortion while preserving local and geometric information [34, 47]. There have been attempts to represent and process mesh and image features in hyperbolic space [22, 39, 38, 1, 2, 23]. However, joint feature learning of meshes and images in hyperbolic space for accurate hand-object reconstruction has not yet been explored.

To this end, we propose the first method based on hyperbolic space for hand-object reconstruction, named Dynamic Hyperbolic Attention Network (DHANet), to leverage the benefits of hyperbolic space for geometrical feature learning (see Fig. 1). Our approach consists of three modules, image-to-mesh estimation, dynamic hyperbolic graph convolution, and image-attention hyperbolic graph convolution. Firstly, the image-to-mesh estimation module geometrically approximates the hand and object from an input image. Secondly, hand and object meshes are projected to hyperbolic space for better preserving the geometrical information. Our dynamic hyperbolic graph convolution dynamically builds neighborhood graphs in hyperbolic space to learn mesh features with rich geometric information. Thirdly, we project mesh and image features to a unified hyperbolic space, preserving the spatial distribution between hand and object. Our image-attention hyperbolic graph convolution embeds the distribution into feature learning and models the hand-object interaction in a learnable way. With these modules, our method learns more representative geometry-image multi-modal features for accurate hand object reconstruction. Comprehensive evaluations of our method on three public hand-object datasets, namely Obman dataset [17], FHB dataset [10], and HO-3d dataset [14], where DHANet outperforms most state-of-the-art methods, confirm the superiority of our design.

The main contributions of our work are as follows:

- We are the first to address hand-object reconstruction in hyperbolic space, proposing a novel Dynamic Hyperbolic Attention Network.
- We devise a Dynamic Hyperbolic Graph Convolution to dynamically learn mesh features with rich geometry information in hyperbolic space.
- We introduce an Image-attention Graph Convolution to learn geometry-image multi-modal features and to model hand-object interactions in hyperbolic space.

2. Related work

2.1. Hand-Object Reconstruction

Hand-object reconstruction is an attractive research area. Earlier methods focused on reconstructing hand and object from multi-view images [4, 32] or RGBD images [13, 20] due to severe occlusion between hand and object. In recent trends, joint reconstruction of both shapes from a single RGB image has become popular. It is a more challenging task due to the limited perspective. Existing methods can be divided into two categories: optimization-based and learning-based methods.

Optimization-based methods design contact patterns manually based on a parameterized representation of hand and object to model the hand-object interaction explicitly. Cao *et al*. [5] leveraged the 2D image cues and 3D contact priors to constrain the optimizations. 2D image cues include the estimated object mask via differentiable rendering and the estimated depth. 3D contact priors are based on hand-object distance and collision. Yang *et al*. [46] presented an explicit contact representation, Contact Potential Field (CPF). Each contacting hand-object vertex pair is treated as a spring-mass system. They also introduced contact constraint items and grasping energy items in their learning-fitting hybrid framework. Ye *et al*. [48] parameterizes the object by signed distance, leveraging the input visual feature and output hand mesh information to infer the object representation. Zhao *et al*. [49] represents hand and object as a hand-object ellipsoid, recovering hand-object driven by the simulated stability criteria in the physics engine. However, the performance of these methods is limited by the manually defined interaction.

Learning-based methods employ advanced mechanisms to model the relationship between hand and object implicitly. These methods can be divided into two categories: non-graph-based and graph-based. **Non-graph-based methods** model without the use of graph structure.The first end-to-end learnable model is presented by Hasson *et al*. [17] that exploits a contact loss to model the interaction. Cheng *et al*. [8] propose a pose dictionary learning module to distinguish infeasible poses. Liu *et al*. [28]

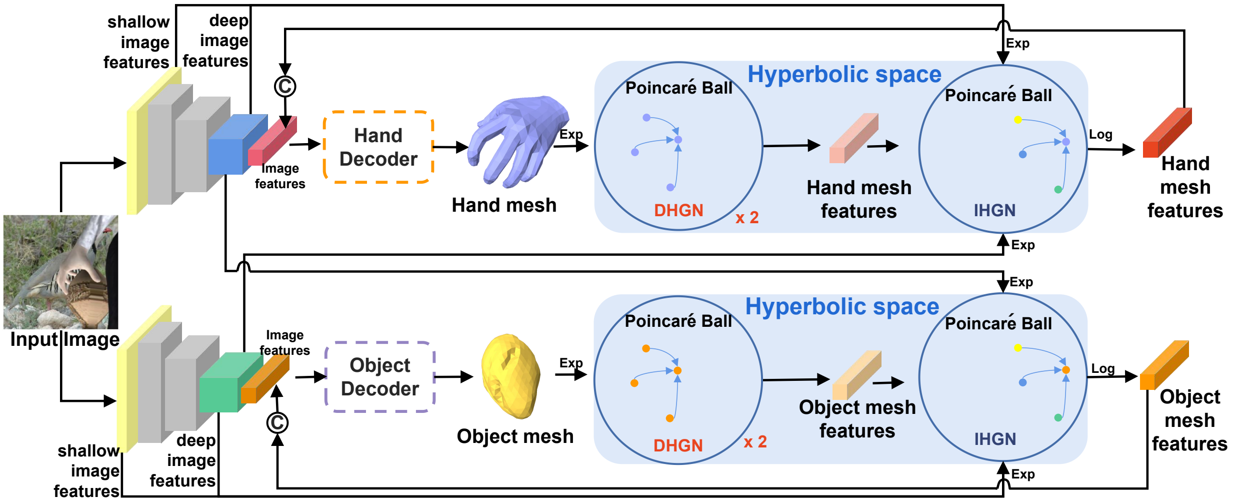

Figure 2. DHANet overview. Given an image with hand-object interaction, image encoder-decoders first approximate the mesh with an initial form. Subsequently, image features from encoders and meshes are projected to hyperbolic space via the *Exp* function. Our dynamic hyperbolic graph convolution (DHGN) and image-attention hyperbolic graph convolution (IHGN) learn representative mesh features, projected to Euclidean space via the *Log* function and concatenated with image features to derive an accurate hand-object reconstruction.

builds a joint learning framework where they performed contextual reasoning between hand and object representations. Li *et al*. [45] propose ArtiBoost, a lightweight online data enhancement method that constructed diverse hand-object interactions using a data enhancement approach. **Graph-based methods** represents hand-object as graphs, utilizing graph convolution to learn the hand-object interaction. Doosti *et al*. [9] are the first to design an Adaptive Graph U-Net to transform 2D keypoints to 3D. A context-aware graph network and a learnable physical affinity loss are proposed to learn interaction messages [51]. Tse *et al*. [42] transfer mesh information to the decoder of image features in a collaborative learning strategy. An attention-guided graph convolution learns mesh information. However, these methods learn the embedding of keypoints or meshes in Euclidean space, failing to capture rich geometry information. In our work, we aim to capture geometry information in hyperbolic space, which are beneficial to the reconstruction of hands and objects.

2.2. Hyperbolic Neural Networks

Recently, incremental works have been done for deep representation learning in hyperbolic spaces [34]. Compared with Euclidean space, hyperbolic space is more suitable for processing data with a tree-like structure or power-law distribution, owing to its exponential growth property [47]. The deep representation learning in hyperbolic space is named hyperbolic neural networks. Existing research mostly focuses on NLP tasks [40, 12, 50, 41]. Recently, some researchers propose to learn hyperbolic embeddings for images in computer vision tasks [24, 33, 3]. It has been confirmed that a similar hierarchical structure exists in images as well. Montanaro *et al*. [31] is the first to apply hyperbolic neural networks to point clouds, demonstrating that a point cloud is a local-whole hierarchical structure. As far as we know, we are the first to seek hand-object reconstruction in hyperbolic space. Hyperbolic space is more suitable for processing meshes than Euclidean space, proven in traditional computer graphics. Researchers project meshes into hyperbolic space to learn parameterized representation in shape analysis and model registration tasks [22, 39, 38, 1]. We embed hand-object meshes into hyperbolic space to learn geometry information.

3. Preliminaries

3.1. Hyperbolic Space

Hyperbolic space is a non-Euclidean space that can be represented as a Riemannian manifold with a constant negative curvature. There are multiple isometric hyperbolic models, including Poincaré ball model, Lorentz model, and Klein model. In this work, we use the Poincaré ball model because it is a conformal geometry, *i.e*. geometry-preserving.

The Poincaré ball model is defined as a Riemannian manifold $\left(B_c^n, g_x^B\right)$, where x is a point in a Riemannian manifold and B represents the Poincaré ball model. $B_c^n = \left\{x \in \mathbb{R}^n : \|x\|^2 < -\frac{1}{c}\right\}$ is an n-dimensional ball with radius $\frac{1}{\sqrt{|c|}}$, and c $(c < 0)$ is the negative curvature of the ball. $g_x^B = (\lambda_x^c)^2 g^E$ is its Riemannian metric, which is conformal to the Euclidean metric g^E with the conformal factor $\lambda_x^c = \frac{2}{1-\|x\|^2}$.

To project points from Euclidean space to hyperbolic space, a mapping function called the exponential mapping function Exp : $\tau_x M \to M$ is defined, as explained in [47]. Here, M refers to the hyperbolic space in the Poincaré ball model, and $\tau_x M$ is its tangent space. The logarithmic map Log is the inverse of Exp and maps points from hyperbolic space to its tangent space, also described in [47]. If we have two points x and y in B_c^n, the distance between them is the geodesic distance $d(x, y)$. This distance is defined as the shortest length of a curve, as further explained in [47].

3.2. Hyperbolic Graph Neural Network

Hyperbolic Graph Neural Networks (HGNN) [27] generalizes graph neural networks to hyperbolic space. In comparison to graph neural networks in Euclidean space, HGNN is more suitable for tree-like data and therefore learns more powerful geometrical representations [47]. HGNN consists of four steps: feature projection, feature transformation, neighborhood aggregation, and activation. For the l-th layer in HGNN, given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with a vertex set $\mathcal{V}$ and an edge set $\mathcal{E}$, $x_i^{l-1,E} \in \mathcal{V}$ is the input node feature for i-th vertex in Euclidean space. The feature projection is to project node features to hyperbolic space by Exp function. The feature transformation is usually operated by a $M\ddot{o}bius$ layer [25], which involves $M\ddot{o}bius$ vector multiplication $\otimes$ and $M\ddot{o}bius$ bias addition $\oplus$. The neighborhood features are aggregated by hyperbolic aggregation functions, AGG^B. The last is a non-linear hyperbolic activation, σ^B. In short, a hyperbolic graph convolution layer can be formulated as:

$$x_i^{l-1,B} = \text{Exp}(x_i^{l-1,E}), \tag{1}$$

$$h_i^{l,B} = x_i^{l-1,B} \otimes W^l \oplus b^l, \tag{2}$$

$$y_i^{l,B} = \text{AGG}^B(h^{l,B}), \tag{3}$$

$$x_i^{l,B} = \sigma^B(y_i^{l,B}). \tag{4}$$

For more details on the functions, please refer to [47].

4. Methodology

In this section, we present our novel method for hand object reconstruction, called the Dynamic Hyperbolic Attention Network (DHANet). As shown in Figure 2, our approach consists of a two-branch network that jointly reconstructs both the hand and object meshes. Specifically, our method comprises three main steps: 1) Image-to-mesh estimation (Section 4.1), 2) Dynamic Hyperbolic Graph Convolution for learning mesh features (Section 4.2), and 3) Image-attention Hyperbolic Graph Convolution for modeling the hand-object interaction (Section 4.3).

4.1. Image-to-mesh estimation

As depicted in Fig. 2, the image-to-mesh estimation step aims to estimate the initial 3D meshes of the hand and object from a given image. Each branch employs an encoder-decoder architecture, where the encoder consists of two pre-trained ResNet-18 [18] encoders on ImageNet [37]. The decoders output the hand and object meshes respectively.

Hand Reconstruction Decoder. The hand reconstruction decoder predicts the hand parameters from image features using the MANO model [36], which is an articulated mesh deformation model rigged with 21 skeleton joints. The MANO model is represented by a differentiable function $D(\beta, \theta)$, where $\theta \in \mathbb{R}^{51}$ denotes the shape parameters and $\beta \in \mathbb{R}^{10}$ denotes the pose parameters. We employ a multi-layer perceptron (MLP) to directly regress β and θ from the image features. Then, a differentiable MANO layer [17] applies D to generate a hand MANO model from β and θ. The hand mesh of the MANO model is defined as $m_h = (v_h, f_h)$, where $v_h \in \mathbb{R}^{778\times3}$ denotes the mesh vertices and $f_h \in \mathbb{R}^{1538\times3}$ denotes the mesh faces. The supervision signal for this branch comes from the L2 loss, which consists of the L2 distance between the predicted mesh vertices and the ground truth mesh vertices, as well as the L2 distance between the predicted joint positions and the ground truth joint positions.

Object Reconstruction Decoder. The objective of the decoder for object reconstruction is to predict the 3D object mesh from the image features. We employ AtlasNet [11] as the object decoder, following the approach of existing methods such as [17, 42, 7]. The AtlasNet branch takes the image features from the encoder and generates the object mesh $m_o = (v_o, f_o)$, where $v_o \in \mathbb{R}^{642\times3}$ represents the mesh vertices and $f_o \in \mathbb{R}^{1280\times3}$ represents the mesh faces. The branch is trained to minimize the Chamfer distance [11], which measures the average minimum distance between points on the predicted mesh and the nearest points on the ground truth mesh.

4.2. Dynamic hyperbolic graph convolution

Hyperbolic space has been shown to be well-suited for processing tree-like graphs due to its exponential growth property, which preserves local and geometric information with low distortion [34, 47]. As meshes are naturally tree-like graphs, we aim to learn mesh features with rich geometric information in hyperbolic space. Inspired by DGCNN [43], which captures the local geometry structure of point clouds in Euclidean space, we propose a dynamic hyperbolic graph convolution to learn mesh features. This module consists of three steps: projection, graph construction, and hyperbolic graph convolution.

Projection. We project the vertices of a mesh $v^E \in \mathbb{R}^{n\times3}$ into hyperbolic space using an exponential map function, $v^B = \text{Exp}(v^E)$, as illustrated in Fig. 3. Here, v^B denotes the set of mesh vertices in hyperbolic space.

Graph Construction. To construct a neighborhood graph for the vertices, we employ a hyperbolic k-nearest

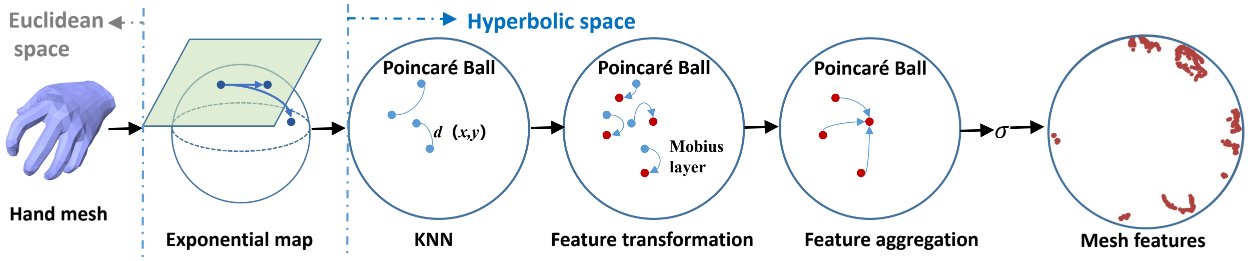

Figure 3. This figure illustrates the pipeline of DHGC, which involves several steps. A given mesh is projected from Euclidean to hyperbolic space using the exponential function. We then conduct dynamic graph construction and employ hyperbolic graph convolution to learn the geometry features of the mesh.

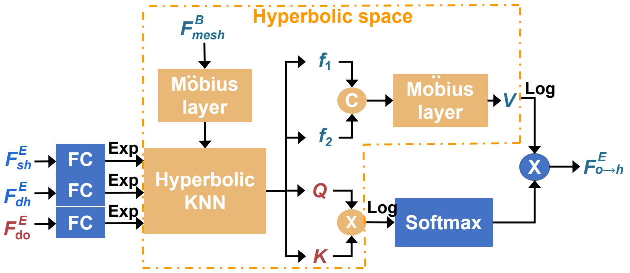

Figure 4. Our image attention hyperbolic graph convolution. The operations in the yellow rectangle are implemented in hyperbolic space, while the blues are in Euclidean space.

neighbors (k-NN) algorithm, which searches for the k closest points for each vertex based on the geodesic distance between two vertices, $d(v_i^B, v_j^B)$. This approach allows us to capture the local geometry structure of the mesh in hyperbolic space.

Hyperbolic Graph Convolution. Hyperbolic graph convolution is to learn a neighborhood feature for each vertex, including transforming vertex features on an m-dimensional Poincaré ball by a $M\ddot{o}bius$ layer [25], aggregating and activating neighborhood features, as shown in Fig. 3. This whole process can be formulated as

$$v^{l,B} = \sigma^B(AGG^B(M\ddot{o}bius(\exp(v^{l-1,E})))). \quad (5)$$

For the aggregation function, we adapt mean aggregation in Poincaré ball, which returns the Einstein midpoint among vertices in a k-neighborhood [47]. Compared to EdgeConv in DGCNN [43], DHGC solely focuses on learning pointwise node features without considering edge features, as there is no defined edge vector in hyperbolic space unlike in Euclidean space.

4.3. Image-attention hyperbolic graph convolution

As mentioned in Section 2.2, due to the exponential growth of distance in hyperbolic space, image features projected to hyperbolic space are more expressive for semantic segmentation [2] and image classification [23]. Inspired by these works, we project image features to hyperbolic space.

Projected image features preserve the spatial relationship between hand and object, which is beneficial for modeling hand-object interaction, as shown in Fig. 1. Hence, we propose an image-attention hyperbolic graph convolution to learn geometry-image multi-modal features, modeling hand-object interaction. As shown in Fig. 4, this module consists of four steps, projection, neighborhood graph construction, feature transformation, and image attention.

Inputs. Taking hand reconstruction as an example, this module takes as input image features in Euclidean space and mesh features in hyperbolic space denoted as F^B_{mesh}. The image features include shallow image features of the hand F^E_{sh}, deep image features of the hand F^E_{dh}, and deep image features of the object F^E_{do}. To ensure consistency in dimensions, fully connected layers are applied to the image features

Projection. We use the exponential function defined in Eq. (1) to map the image features to hyperbolic space. This results in obtaining image features in hyperbolic space denoted as F^B_{sh}, F^B_{dh}, F^B_{do}.

Graph Construction. To construct the k-neighborhood for each vertex in F^B_{mesh}, we utilize a hyperbolic KNN algorithm Specifically, We construct four types of k-neighborhood for each vertex. These four neighborhoods of each vertex are successively composed of k mesh features, k shallow image features, k deep hand image features and k deep object image features, which are defined as f_1, f_2, Q and K. Through building four types of k-neighborhood, image features, and mesh features are aligned in a unified hyperbolic space.

Feature Transformation. Mesh features are enhanced by similar shallow image features. In a neighborhood, mesh features f_1 are concatenated with similar shallow image features f_2. The concatenated feature is transformed into V with a similar dimension as Q and K, by a $M\ddot{o}bius$ layer, formulated as:

$$V = M\ddot{o}bius(Cat(f_1, f_2)), \quad (6)$$

where Cat represents the concatenation operation.

Image Attention. We define the image attention to model the hand-object interaction. V indicates hand mesh features. Q refers to deep image features of hands, which are similar to hand mesh, while K refers to deep image features of objects, which are similar to hand mesh. Then we use the object image feature to fetch the hand image feature and hand mesh feature, as shown Fig. 4. The process can be formulated as

$$F^E_{o\to h} = \mathrm{softmax}(\frac{\mathrm{Log}(Q)\mathrm{Log}(K)^T}{\sqrt{d}})\mathrm{Log}(V), \qquad (7)$$

where $F^E_{o\to h}$ is the hand-object attention mesh features encoding the interaction between hand and object, and d is a normalization constant. For ease of calculation, we map features by Log function into Euclidean space. At last, image-attention hyperbolic graph convolution learns geometry-image multi-modal features, concatenated with image features from encoders to reconstruct a mesh by decoders, as shown in Fig. 2.

5. Implement Details of DHANet

Architecture. Given an input image I with size 256×256, DHANet reconstructs a hand mesh m_h of size 778×3, and an object mesh m_o of size 642×3. Our DHANet is a two-branch network, one for hand reconstruction and the other one for object reconstruction. In the hand branch, the encoder, ResNet-18 [19], extracts 512-dimensional image features F_h, shallow image features F^E_{sh} with size $64 \times 64 \times 32$, and deep image features F^E_{dh} with size $8 \times 8 \times 64$. Then the decoder consisted of fully connected layers and a MANO layer initially reconstructs a hand mesh m_h from F_h. With the hand mesh as input, two stacked DHGC layers learn mesh features $F^B_{h,mesh}$ of size 778×64 in hyperbolic space. The IHGC learns enhanced 32-dimensional mesh features, $F^E_{o\to h}$, inputting $F^B_{h,mesh}$, F^E_{sh}, F^E_{dh} and deep image features of objects F^E_{do}. At last, mesh features are concatenated with 512-dimensional image features, reconstructing hand and object mesh by the decoder. The architecture of the object branch is similar to the hand branch. In order to reconstruct the photo-consistent hand and object, we also adopt two fully connected layers to estimate the 3D offset coordinates T for the translation and a scalar S for the scale, as in [17]. For more details on the detailed architecture of our DHANet please refer to the supplementary material.

Loss function. To supervise the training of our DHANet, we use the loss items in Hasson *et al.* [17], except for the contact loss term, defined as:

$$\mathcal{L} = \mathcal{L}_{hand} + \mathcal{L}_{obj} \qquad (8)$$

$$\mathcal{L}_{hand} = \mathcal{L}_{V_{hand}} + \mathcal{L}_J + \mathcal{L}_\beta \qquad (9)$$

$$\mathcal{L}_{obj} = \mathcal{L}_{V_{obj}} + \mathcal{L}_T + \mathcal{L}_S. \qquad (10)$$

In Eq. (8), $\mathcal{L}_{hand}$ for the hand branch consists of the L2 loss of vertex positions $\mathcal{L}_{V_{hand}}$, the L2 loss of hand joints $\mathcal{L}_J$ and the L2 loss of the hand shape $\mathcal{L}_\beta$. $\mathcal{L}_{obj}$ for the object branch includes the Chamefer distance $\mathcal{L}_{V_{obj}}$, the L2 loss of object scale $\mathcal{L}_S$ and the L2 loss of object translation $\mathcal{L}_T$. $\mathcal{L}_T$ is defined as $\mathcal{L}_T = \left\|T - \hat{T}\right\|_2^2$, where $\hat{T}$ is the ground truth object centroid in hand-relative coordinates. $\mathcal{L}_S$ is defined as $\mathcal{L}_S = \left\|S - \hat{S}\right\|_2^2$, where $\hat{S}$ is the ground truth maximum radius of the centroid-centered object.

6. Experiments

6.1. Datasets

Obman is a large-scale synthetic image dataset of hands grasping objects [17]. The objects in Obman are 8 types of common items, whose models are selected from the ShapeNet [6] dataset. The hands in this dataset are modeled with MANO [35]. The dataset is labeled with 3D hand and object meshes, divided into 141K training frames and 6K test frames.

First-person hand benchmark (FHB) is a real egocentric RGB-D videos dataset about hand-object interaction [10]. There are 105,459 RGB-D frames annotated with 3D object meshes for 4 items and the 3D location of hand joints. We use the same way to divide a training set and a testing set, like [17]. To be consistent with the existing methods [17, 42], we exclude the milk model and filter frames in which the hand is further than 10 mm from the object. This subset of FHB is called FHB$^-$.

HO-3D is also a real image dataset for hand-object interaction [14]. The objects in HO-3D are 10 objects from YCB dataset [44]. The dataset contains hand-object 3D pose annotated RGB images and their corresponding depth maps. Our experiment uses HO-3D (version 2) split into 70K training images and 10K evaluation images as in [16].

6.2. Evaluation metrics

The reconstruction quality of hands and objects are evaluated with the following metrics.

Hand error. The mean end-point error (mm) over 21 joints and the mean vertices error of meshes are computed to evaluate the hand reconstruction.

Object error. To evaluate the object reconstruction, we report the Chamfer distance (mm) between points sampled on the ground truth mesh and vertices of the predicted mesh.

Contact metrics. Reconstructed hand-object should be impenetrable according to the laws of physics. For assessing the physical validity of the results, we also adopt penetration depth (mm) and intersection volume (cm^3) as [17, 42]. Penetration depth is the maximum distance between hand mesh and object mesh when the hand collides with the object. Otherwise, the penetration depth is 0. Intersection volume is the volume of the interaction area btween

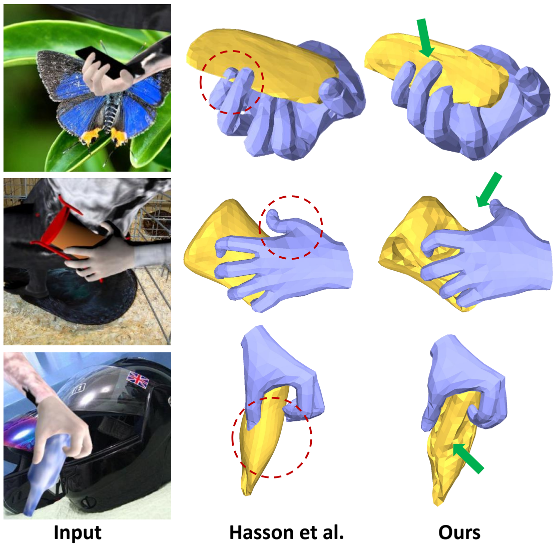

Figure 5. Qualitative comparison with Hasson *et al.* [17] on Obman dataset [17]. The red circles highlight the errors from Hasson *et al.* [17]. The green arrows point to improvements of our method.

Methods	Hand error	Object error	Max. penetra.	Intersect. vol.
Hasson *et al.* [17]	11.6	641.5	9.5	12.3
Tse *et al.* [42]	**9.1**	**385.7**	**7.4**	**9.3**
Ours	10.2	529.3	9.3	10.4

Table 1. Comparison to state-of-the-art methods on Obman dataset [17]. The hand error is calculated on joints. Here we report the result of Hasson *et al.* [17] without contact loss. "Max penetration" is shortened to "Max. penetra.". "Intersection volume" is shortened to "Intersect. vol."

the hand and object. We compute the volume by voxelizing the hand and object under a voxel size of 0.5 cm.

6.3. Training details

We adopt the work of Hasson *et al.* [17] as the backbone to pre-estimate rough hand-object meshes, namely baseline. The model parameters of the baseline are initialized by the pre-trained model of Hasson *et al.* [17]. We choose the Riemannian Adam optimizer to train our DHANet, since [25] verified that the Riemmanian Adam optimizer speeds up model convergence for hyperbolic space and works as the standard Adam optimizer for Euclidean space. The training for different datasets is different. For the Obman dataset, the training strategy is the same as [17]. We first train the object branch for 100 epochs at a learning rate 10^{-4}, then train the hand branch for 100 epochs at a learning rate 10^{-4} while freezing the object branch. For datasets of real scenes, HO-3d and FHB^{-}, we train the hand and object branches together for 300 epochs with a learning rate of 10^{-4}, then train them for other 300 epochs with a learning rate of 10^{-5}.

Methods	Hand error	Object error	Max. penetra.	Intersect. vol.
Hasson *et al.* [17]	28.1	1579.2	18.7	26.9
Tse *et al.* [42]	25.3	1445.0	16.1	**14.7**
Ours	**23.8**	**1236.0**	**14.43**	20.7

Table 2. Comparison to state-of-the-art methods on FHB^{-} dataset [10]. The hand error is calculated on joints. Here we report the result of Hasson *et al.* [17] without contact loss. "Max penetration" is shortened to "Max. penetra.". "Intersection volume" is shortened to "Intersect. vol."

Methods	Hand error	Object error
Hasson *et al.* [16]	14.7	26.8
Cao *et al.* [5]	9.7	19.9
Tse *et al.* [42]	10.9	-
Ours	**6.1**	**13.8**

Table 3. Comparison to state-of-the-art methods on HO-3d dataset [14]. The hand error is calculated on the vertices of the hand mesh.

6.4. Hand-object reconstruction results

Method for comparison. In the single image hand-object reconstruction field, there are a few methods [17, 5, 46, 42, 26], which represent a hand as MANO model and represent an object as 3D mesh. While existing methods include two categories, one for known object models [46, 5, 26], the other for unknown object models [17, 42], Our method belongs to the latter category. There are still some hand-object reconstruction works based on SDF [7, 48], representing objects as a dense 3D mesh, while in our work we reconstruct an object as a simple mesh with 642 vertices. Hence, we compare our method with [17] and [42].

Results. Table 1 indicates our method achieves better results on Obman dataset [17] than the baseline method [17] in hand and object errors. Compared with the baseline method [17], our method yields a smaller hand error of 10.7 mm vs. 11.6 mm and a smaller object error of 563.5 mm vs. 641.5 mm. Our method also achieves better results on contact metrics. As shown in Fig. 5, our method reconstructs better the fine-grained pose and shape of hands with respect to the input image. Like the drum in Fig. 5, the reconstructed drum by our method is more consistent with the original shape in the image. And it can be observed that hands reconstructed by our method are penetrated less with objects. This suggests that our method better models hand-object interaction. However, the performance of our method is less than Tse *et al.* [42] in Table 1. The reason is that the work of Tse *et al.* [42] is a dual-iterative network, in contrast to our DHANet that operates without iteration. While the two iterations of Tse *et al.* [42] yield a good result, they

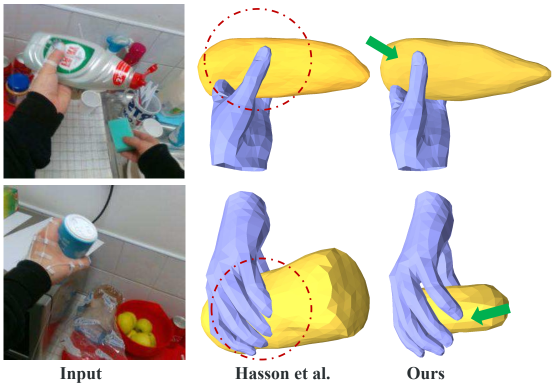

Figure 6. Qualitative comparison with Hasson *et al*. [17] on FHB$^-$ dataset [10].The red circles hightlight the errors from Hasson *et al*. [17]. The green arrows point to improvements of our method.

increase the model parameter simultaneously.

The experimental results compared with existing methods on FHB$^-$ dataset [10] is listed in Table 2. In FHB$^-$ dataset, our method achieves SOTA results whit smaller hand error (23.8 mm) and smaller object error (1236.0 mm). The qualitative results of this dataset are shown in Fig. 6. Our method exceeds the work of Tse *et al*. [42] on FHB$^-$ dataset in Table 2, but slightly worse on Obman dataset in Table 1. The reason for this result is dataset difference: Obman dataset is a synthetic dataset with complex and various backgrounds, while FHB$^-$ dataset is a real-world dataset with a simplistic kitchen environment background. This results in shallow image features learned on Obman containing more irrelevant background information. The more irrelevant features are, the more loose their projections are in hyperbolic space, and vice versa. We drew the visualization of it in the supplemental material. Those loose shallow image features in hyperbolic space are unfavorable to searching the neighborhood between those features and mesh features. This causes our approach to work slightly worse in Obman dataset.

And the comparison results on HO-3d [14] are shown in Table 3. We also reach SOTA results on the hand error and the object error. FHB$^-$ dataset and HO-3d are captured in real scenes, not synthetic data. The decent results manifest our method can handle not only synthetic data but also real-world cases.

6.5. Ablation study

We conducted ablation studies to demonstrate the effectiveness of our proposed dynamic hyperbolic graph convolution (DHGC) and image-attention hyperbolic graph convolution (IHGC). As shown in Table 4, adding DHGC with baseline reduces the hand error to 10.9 mm while reducing the object error to 582.9 mm. This suggests that the mesh feature learned by DHGC provides richer geometric information. Furthermore, IHGC further improves the reconstruction results, which further reduced the hand and object error to 10.2 mm and 529.3 mm. And the performance on contact metrics also declined. These results demonstrate that IHGC effectively enhances mesh features with image features while modeling hand-object interactions.

In order to verify the superiority of our method in hyperbolic space, we also implement dynamic graph convolution and image-attention graph convolution in Euclidean space. The comparison results are enumerated in Table 4. We can observe that these two modules in Euclidean space have improved from the baseline [17], while the improvement is less than ours in hyperbolic space. It proves quantitatively that our method achieves better performance in hyperbolic space than in Euclidean space.

Methods	Hand error	Object error	Max. penetra.	Intersect. vol.
Baseline [17]	11.6	641.5	9.5	12.3
Baseline+1(EU)	11.6	589.2	10.8	13.5
Baseline+1+2(EU)	11.1	586.1	11.2	10.7
Baseline+1(H)	10.9	582.9	10.7	10.9
Baseline+1+2(H)	**10.2**	**529.3**	**9.3**	**10.4**

Table 4. Ablations on modules and feature spaces. 1 refers to dynamic hyperbolic graph convolution. 2 refers to image-attention hyperbolic graph convolution. EU represents the operation in Euclidean space, while H represents hyperbolic space.

6.6. Visual analysis of hyperbolic learning

We further prove the superiority of our method by visual analysis of features in hyperbolic space, as shown in Fig. 7. To facilitate observation, we use UMAP [29] to project features to 2 dimensions, as in [31]. Given an image as in Fig. 7 (a), we visualized the distribution of the corresponding 3D mesh in hyperbolic space and Euclidean space, depicted in Fig. 7 (f) and Fig. 7 (g). As shown in Fig. 7 (c), pink points are near blue points and far from brown points. The relative position is reflected in hyperbolic space, as depicted in Fig. 7 (f), but not in Euclidean space, as depicted in (g). It indicates that embedding mesh into hyperbolic space can preserve the geometry properties of the mesh.

In image-attention hyperbolic graph convolution, we project mesh features, shallow image features, deep image features of hand, and deep image features of object to hyperbolic space. In Fig. 7 (h), some yellow points are overlapping with red points. It indicates that shallow image features are aligned with mesh features in hyperbolic space. However, shallow image features and mesh features are separated in Euclidean space, as shown in Fig. 7 (i). Aligned features are conducive to feature learning in a unified space. In addition, there are overlapping regions in deep image features of hand and object, as shown in Fig. 7 (d) and Fig. 7 (e), expressing the area of hand-object interaction. It is re-

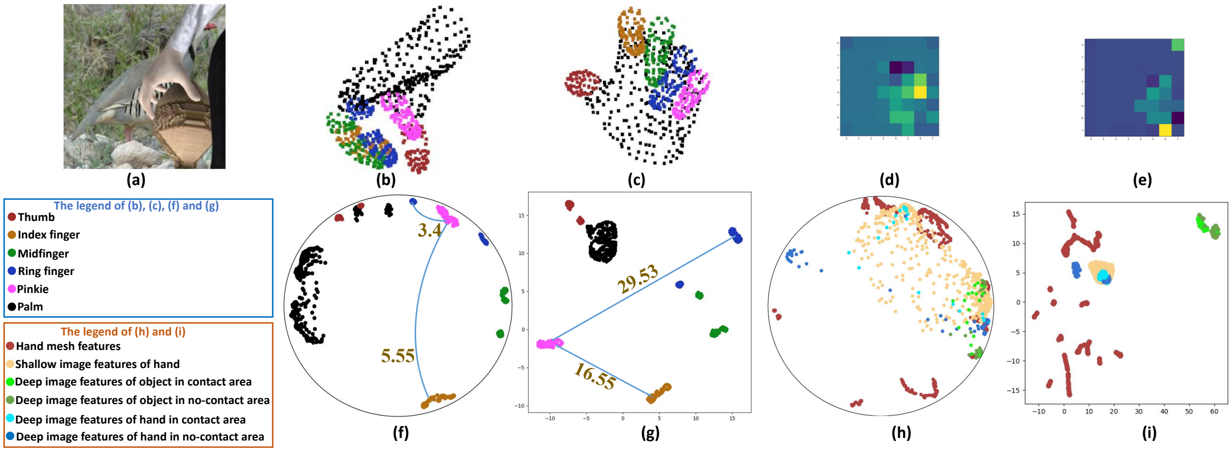

Figure 7. Visualization of features in hyperbolic space and Euclidean space. (a): a sample image from Obman dataset [17]. (b): vertices of the hand mesh reconstructed from (a). (c) is rotated by (b). (d): the hand deep image features from the encoder of the hand branch. (e): the object deep image features from the encoder of the object branch. The description of (f), (g), (h), (i), and (j) is in Section 6.6.

flected in hyperbolic space, as shown in Fig. 7 (h). Some light blue points are close to a few light green points, others vice versa. The closer region in hyperbolic space represents the area of hand-object interaction in the image. Furthermore, the spatial relationship is not expressed in Euclidean space. As shown in Fig. 7 (i), the light blue points are far from the light green points. This highlights the ability of hyperbolic space to align multi-modal features and preserve spatial relationships.

7. Conclusion

In this work, we propose a dynamic hyperbolic graph neural network (DHANet) for hand object reconstruction. Our method applies hyperbolic neural networks for the first time in this task. By leveraging hyperbolic space, we design a dynamic hyperbolic graph convolution that captures rich geometry information in mesh features. By projecting multi-modal features to a unified hyperbolic space, we define a more accurate representation of geometry-image features. To model the hand-object interaction, we introduce an attention-based hyperbolic graph convolution that enhances mesh features with image features. Our method outperforms state-of-the-art methods on public datasets, achieving more accurate reconstruction of hand and object meshes. This approach offers a new perspective for hand object reconstruction in hyperbolic space, with promising opportunities for future research. As for future work, since image features are vital for modeling interaction and affect our performance, we plan to explore compositional image features, such as extracted from hand or object parts.

Acknowledgment

This work was supported by the National Nature Science Foundation of China under Grant 62272019, in part by China Scholarship Council.

References

[1] Noam Aigerman and Yaron Lipman. Hyperbolic orbifold tutte embeddings. *ACM Trans. Graph.*, 35(6):217–1, 2016. 2, 3

[2] Mina Ghadimi Atigh, Julian Schoep, Erman Acar, Nanne Van Noord, and Pascal Mettes. Hyperbolic image segmentation. In *CVPR*, pages 4453–4462, 2022. 2, 5

[3] Mina Ghadimi Atigh, Julian Schoep, Erman Acar, Nanne Van Noord, and Pascal Mettes. Hyperbolic image segmentation. In *CVPR*, pages 4453–4462, 2022. 3

[4] Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using discriminative salient points. In *ECCV*, pages 640–653. Springer, 2012. 2

[5] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *ICCV*, pages 12417–12426, 2021. 1, 2, 7

[6] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 6

[7] Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In *ECCV*, pages 231–248. Springer, 2022. 4, 7

[8] Zida Cheng, Siheng Chen, and Ya Zhang. Semi-supervised 3d hand-object pose estimation via pose dictionary learning. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3632–3636. IEEE, 2021. 2

[9] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *CVPR*, pages 6608–6617, 2020. 1, 2, 3

[10] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action bench-

mark with rgb-d videos and 3d hand pose annotations. In *CVPR*, pages 409–419, 2018. 2, 6, 7, 8

[11] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *CVPR*, pages 216–224, 2018. 4

[12] Caglar Gulcehre, Misha Denil, Mateusz Malinowski, Ali Razavi, Razvan Pascanu, Karl Moritz Hermann, Peter Battaglia, Victor Bapst, David Raposo, Adam Santoro, et al. Hyperbolic attention networks. *arXiv preprint arXiv:1805.09786*, 2018. 3

[13] Henning Hamer, Juergen Gall, Thibaut Weise, and Luc Van Gool. An object-dependent hand pose prior from sparse training data. In *CVPR*, pages 671–678. IEEE, 2010. 2

[14] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, pages 3196–3206, 2020. 1, 2, 6, 7, 8

[15] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *CVPR*, pages 11090–11100, 2022. 1, 2

[16] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *CVPR*, pages 571–580, 2020. 6, 7

[17] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, pages 11807–11816, 2019. 1, 2, 4, 6, 7, 8, 9

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4

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

[20] Haoyu Hu, Xinyu Yi, Hao Zhang, Jun-Hai Yong, and Feng Xu. Physical interaction: Reconstructing hand-object interactions with physics. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 2

[21] Lin Huang, Jianchao Tan, Jingjing Meng, Ji Liu, and Junsong Yuan. Hot-net: Non-autoregressive transformer for 3d hand-object pose estimation. In *ACM International Conference on Multimedia*, pages 3136–3145, 2020. 1, 2

[22] Miao Jin, Feng Luo, and Xianfeng Gu. Computing surface hyperbolic structure and real projective structure. In *Proceedings of the 2006 ACM symposium on Solid and physical modeling*, pages 105–116, 2006. 2, 3

[23] Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *CVPR*, pages 6418–6428, 2020. 2, 5

[24] Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *CVPR*, pages 6418–6428, 2020. 3

[25] Max Kochurov, Rasul Karimov, and Serge Kozlukov. Geoopt: Riemannian optimization in pytorch. *arXiv preprint arXiv:2005.02819*, 2020. 4, 5, 7

[26] Kailin Li, Lixin Yang, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. *arXiv preprint arXiv:2109.05488*, 2021. 7

[27] Qi Liu, Maximilian Nickel, and Douwe Kiela. Hyperbolic graph neural networks. *Advances in neural information processing systems*, 32, 2019. 4

[28] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *CVPR*, pages 14687–14697, 2021. 1, 2

[29] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 8

[30] Antonio Montanaro, Diego Valsesia, and Enrico Magli. Rethinking the compositionality of point clouds through regularization in the hyperbolic space. *arXiv preprint arXiv:2209.10318*, 2022. 2

[31] Antonio Montanaro, Diego Valsesia, and Enrico Magli. Rethinking the compositionality of point clouds through regularization in the hyperbolic space. *arXiv preprint arXiv:2209.10318*, 2022. 3, 8

[32] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *ICCV*, pages 2088–2095. IEEE, 2011. 2

[33] Wei Peng, Jingang Shi, Zhaoqiang Xia, and Guoying Zhao. Mix dimension in poincaré geometry for 3d skeleton-based action recognition. In *ACM International Conference on Multimedia*, pages 1432–1440, 2020. 3

[34] Wei Peng, Tuomas Varanka, Abdelrahman Mostafa, Henglin Shi, and Guoying Zhao. Hyperbolic deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10023–10044, 2021. 2, 3, 4

[35] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (SIGGRAPH Asia)*, 36(6), Nov. 2017. 6

[36] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 4

[37] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 4

[38] Jie Shi, Wen Zhang, and Yalin Wang. Shape analysis with hyperbolic wasserstein distance. In *CVPR*, pages 5051–5061, 2016. 2, 3

[39] Rui Shi, Wei Zeng, Zhengyu Su, Hanna Damasio, Zhonglin Lu, Yalin Wang, Shing-Tung Yau, and Xianfeng Gu. Hyperbolic harmonic mapping for constrained brain surface registration. In *CVPR*, pages 2531–2538, 2013. 2, 3

[40] Ryohei Shimizu, Yusuke Mukuta, and Tatsuya Harada. Hyperbolic neural networks++. *arXiv preprint arXiv:2006.08210*, 2020. 3

[41] Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. Hyperbolic representation learning for fast and efficient neural question answering. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 583–591, 2018. 3

[42] Tze Ho Elden Tse, Kwang In Kim, Ales Leonardis, and Hyung Jin Chang. Collaborative learning for hand and object reconstruction with attention-guided graph convolution. In *CVPR*, pages 1664–1674, 2022. 1, 2, 3, 4, 6, 7, 8

[43] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (TOG)*, 38(5):1–12, 2019. 4, 5

[44] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 6

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

[46] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *ICCV*, pages 11097–11106, 2021. 1, 2, 7

[47] Menglin Yang, Min Zhou, Zhihao Li, Jiahong Liu, Lujia Pan, Hui Xiong, and Irwin King. Hyperbolic graph neural networks: A review of methods and applications. *arXiv preprint arXiv:2202.13852*, 2022. 2, 3, 4, 5

[48] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *CVPR*, pages 3895–3905, 2022. 2, 7

[49] Zimeng Zhao, Binghui Zuo, Wei Xie, and Yangang Wang. Stability-driven contact reconstruction from monocular color images. In *CVPR*, pages 1643–1653, 2022. 2

[50] Yudong Zhu, Di Zhou, Jinghui Xiao, Xin Jiang, Xiao Chen, and Qun Liu. Hypertext: Endowing fasttext with hyperbolic geometry. *arXiv preprint arXiv:2010.16143*, 2020. 3

[51] Nan Zhuang and Yadong Mu. Joint hand-object pose estimation with differentiably-learned physical contact point analysis. In *International Conference on Multimedia Retrieval*, pages 420–428, 2021. 1, 2, 3