






3D Keypoint Estimation Using Implicit Representation Learning

Xiangyu Zhu^{1†} , Dong Du^{1,3†} , Haibin Huang⁴ , Chongyang Ma⁴ , Xiaoguang Han^{1,2‡} ¹SSE, CUHKSZ ²FNii, CUHKSZ ³City University of Hong Kong ⁴Kuaishou Technology

Abstract

In this paper, we tackle the challenging problem of 3D keypoint estimation of general objects using a novel implicit representation. Previous works have demonstrated promising results for keypoint prediction through direct coordinate regression or heatmap-based inference. However, these methods are commonly studied for specific subjects, such as human bodies and faces, which possess fixed keypoint structures. They also suffer in several practical scenarios where explicit or complete geometry is not given, including images and partial point clouds. Inspired by the recent success of advanced implicit representation in reconstruction tasks, we explore the idea of using an implicit field to represent keypoints. Specifically, our key idea is employing spheres to represent 3D keypoints, thereby enabling the learnability of the corresponding signed distance field. Explicit keypoints can be extracted subsequently by our algorithm based on the Hough transform. Quantitative and qualitative evaluations also show the superiority of our representation in terms of prediction accuracy.

CCS Concepts

• **Computing methodologies** → **Shape analysis; Shape representations;**

1. Introduction

In this paper, we study the challenging and under-explored problem of 3D keypoint estimation for general shapes. As a key component in many downstream tasks, an accurate and robust 3D keypoint estimation method can provide useful clues for various applications, including 3D object detection [MBO06, LWT20, BABM19], object tracking [SGG*08, CC10, BKB18], shape matching [ZHDQ08, BKB18, WGY*18], and shape registration [BF08, LHM*15, BMS-GJL16].

Although existing methods have demonstrated great success in the detection of facial landmarks as well as human body joints [BADDB11, CPA11, PZK*17, PZK*17, GSX*21], they are designed for shapes with consistent structures. It is commonly not easy to extend such methods for 3D keypoint estimation of general shapes which usually present diverse geometric topologies and irregular numbers of keypoints. Recently, You et al. [YLL*20b] proposed the first large-scale 3D keypoint dataset of 16 general object categories in ShapeNet [CFG*15] and established a benchmark for the task of keypoint prediction. All the methods evaluated in [YLL*20b] focus on complete point cloud input, where the keypoint prediction can be converted into a classification task for each point. However, explicit or complete geometry is typically expensive to obtain. For example, the input for keypoint estimation can

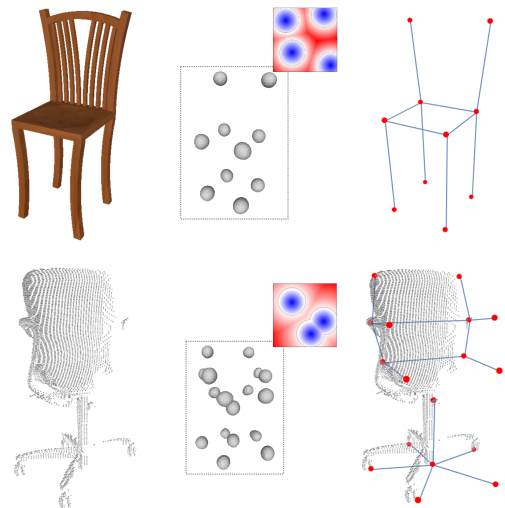


Figure 1: 3D keypoint estimation via our implicit sphere learning. Given a single image or partial point cloud, we learn the SDF of keypoint spheres, and the sphere meshes are extracted for the final keypoint estimation. To enhance visualization, we draw lines connecting the keypoints.

be images or partial point clouds. This makes classification-based methods infeasible since the expected 3D keypoints cannot be explicitly obtained from the input.

[†] Equal contribution.

[‡] Corresponding author: hanxiaoguang@cuhk.edu.cn

To tackle the challenge of keypoint estimation for general objects, alternative methods leverage deep neural networks and can be roughly grouped into two categories, i.e., point coordinate regression [FSG17] and heatmap inference [ORL18]. Methods that directly regress spatial coordinates of keypoints are straightforward but increase the risk of overfitting. Moreover, both order and number of keypoints generally need to be fixed for the design and implementation of the network, which is unreasonable for shapes with varying structures and topologies. On the other hand, the heatmap representation is often proposed for 2D keypoint estimation. The consumption of calculation and storage increases significantly for 3D scenarios, leading to low-resolution heatmap prediction and poor accuracy of keypoint estimation. Neither point regression nor heatmap inference is adequate to generate accurate 3D keypoints in irregular and unordered cases.

Inspired by the recent success of accurate 3D reconstruction with implicit shape learning [MON*19, PFS*19, CZ19, CAPM20], we propose a novel implicit representation for 3D keypoints estimation. Specifically, 3D keypoints are represented as the centers of distinct spheres with a user-specified radius, then the signed distance field (SDF) of these spherical shapes can be inferred using classical implicit learning methods. Given a point cloud or a single-view image, we adopt a deep neural network to learn the SDF field and extract explicit sphere meshes, followed by keypoint extraction using the Hough transform. With this new formulation, we are able to not only handle uncertain number and order properties of general object keypoints but also to improve the performance of keypoint estimation for incomplete point cloud or image inputs, as shown in Figure 1. Furthermore, we explore the semantic keypoint prediction with implicit learning using the proposed stacked unsigned distance field (UDF), in order to benefit applications that require semantic information.

We conduct experiments about 3D keypoint regression on the KeypointNet dataset [YLL*20b] to compare our method with two alternative representations, i.e., the point coordinate and heatmap. Comparisons for complete point cloud, partial point cloud, and single-view image input settings are presented respectively. Both quantitative and qualitative results demonstrate the superiority of our implicit representation for 3D keypoint estimation.

Our contributions can be summarized as follows:

- We introduce the continuous implicit field as a sparse point representation for the first time and propose a consistent 3D keypoint estimation framework for general objects with various topologies and geometry.
- With our implicit representation, we also propose a novel architecture that can generate keypoints with semantic labels.
- We conduct extensive experiments on 3D keypoint estimation with various inputs including complete point clouds, partial point clouds, and single-view images, which demonstrate the superiority of our implicit representation.

In the following sections of this paper, we will first review related work about 3D keypoint detection and estimation, as well as implicit representation learning in Section 2. Then, we propose our implicit keypoint representation and architecture for keypoint learning, extraction, and semantic prediction in Section 3. Next, we evaluate and compare our method with existing approaches both

quantitatively and qualitatively in Section 4. Lastly, we summarize our method and discuss its limitations and future work in Section 5.

2. Related Work

Keypoint detection. 3D keypoint saliency detection, which picks up keypoints from a full point cloud, has been a classical task for many downstream applications, such as object detection, pose estimation, shape matching, and registration. Traditional methods mainly utilize hand-crafted geometric features to select the most salient keypoints, but they either ignore the semantic information of keypoints or tend to generate misaligned keypoints [NN07, Zho09, SOG09, TSDS10, SB11, KZK17]. Li et al. [LL19] pioneer a learning-based 3D keypoint detector, named USIP. However, USIP takes advantage of probabilistic Chamfer loss which may greatly enhance the repeatability of inferred keypoints. Whereafter, Wei et al. [WMW*21] attempt to jointly learn the 3D keypoint saliency and correspondence to improve accuracy. Recently, Fernandez et al. [FLCP*20] propose an unsupervised method to learn aligned 3D keypoints by decomposing keypoint coordinates into low-rank non-rigid shape registration. This approach is suitable for the detection of similar shapes but cannot perform well on general objects with various topologies and geometry. Shi et al. [SXYL21] improve the unsupervised detector with the guidance of skeletons and a proposed composite Chamfer distance. In contrast to these detection-based methods, our method focuses on keypoint generation of general objects, where the inputs can be incomplete (e.g., partial point clouds and single-view images).

Keypoint estimation. Although 3D keypoint detection has achieved great success, it is not suitable for acquiring full keypoints from incomplete inputs. Most of the 3D keypoint regression methods are designed for specific object categories with consistent topologies, such as human faces [EMXD19] and human bodies [KKA19, CYW*19, DZ19, DGM*19, YWS*21]. The keypoint generation of general objects remains challenging since there are diverse topologies and geometric structures in general objects. Recently, He et al. [HSH*20] introduce a voting network for 3D keypoint estimation for point cloud input. Zhou et al. [ZKG*18] propose an unsupervised domain adaptation method for 3D keypoint prediction from a single depth scan or image. Suwajanakorn et al. [SSTN18] also explore an end-to-end geometric reasoning method for the discovery of latent 3D keypoints without supervision. However, this unsupervised method takes pose estimation as a downstream target, and it may not generate useful keypoints for other applications like shape deformation. Vasconcelos et al. [VMB*19] also utilize the domain knowledge for keypoint estimation of general objects, but the performance suffers from the limited dataset. Afterward, You et al. [YLL*20b] provide the first large-scale dataset of annotated keypoints for 16 general object categories. In this paper, we utilize this dataset and propose a unified architecture for 3D keypoint estimation of general objects.

Implicit representation learning. There are various representations for 3D shape learning, such as volumes [CXG*16], point clouds [FSG17], and implicit fields [MON*19, PFS*19, CAPM20]. However, it is not effective or proper to represent 3D keypoint as volumes. Besides, directly regressing 3D point coordinates is not reasonable since it requires a fixed number of points. Inspired by

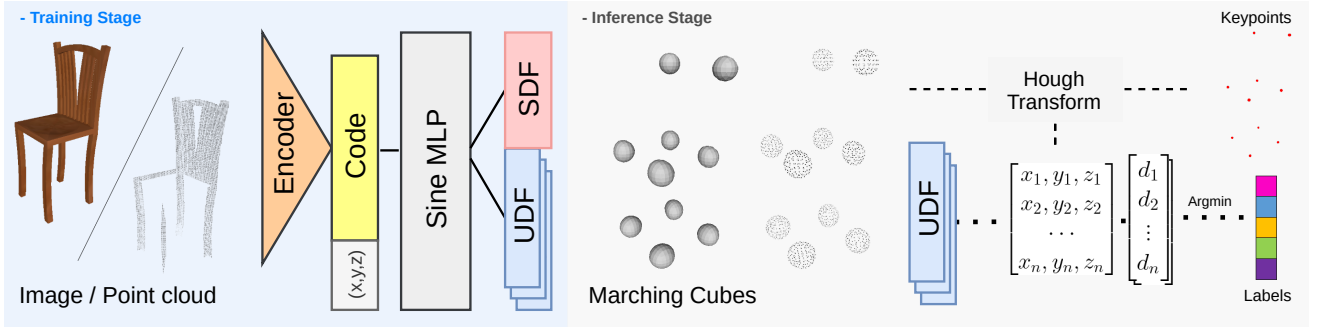


Figure 2: Overview of our implicit keypoint estimation framework. In the inference stage, we extract keypoints from the learned SDF and fetch their coordinates to generate the corresponding semantic labels by the learned stacked UDF.

2D heatmap regression [PCZ15,BT16,ORL18], an alternative keypoint representation is heatmap [PZDD17], but the consumption of calculation and storage increases extremely for 3D scenarios, leading to low-resolution 3D heatmap prediction and poor accuracy of keypoint estimation. As implicit learning has exhibited its great power for 3D reconstruction, we adopt it for 3D keypoint learning in this paper and validate its superiority in our experiments.

3. Method

In this section, we formally introduce our method, including an implicit representation of 3D keypoints and a framework to extract 3D keypoints from various forms of inputs based on the proposed representation. Specifically, we formulate 3D keypoints as implicit spheres represented by SDF and train a deep neural network to infer SDF from various inputs, such as complete/partial point clouds and single-view images (Section 3.1). We then utilize Marching Cubes [LC87] to obtain explicit spheres' surfaces and estimate centers of spheres with Hough transform algorithm [CVC14] as our final keypoints (Section 3.2). Furthermore, we introduce stacked UDF learning for semantic keypoint prediction (Section 3.3).

3.1. Implicit Keypoint Estimation

3.1.1. Implicit Keypoint Network

Implicit representation such as SDF has demonstrated its advantages in shape reconstruction with various topologies and geometric structures [MON*19,PFS*19,CZ19,CAPM20]. In this work, we introduce SDF for 3D keypoint representation to handle irregular and unordered keypoints of general objects. Specifically, we expand a keypoint \mathbf{p}_i to a *keypoint sphere* $\partial B(\mathbf{p}_i, r)$ with a user-specified radius r (r is empirically set as 0.08 and is fixed in all of our experiments) where $B(\mathbf{p}_i, r) = \{x \in \mathbb{R}^3 \mid \|x - \mathbf{p}_i\|_2 \leq r\}$, and define *keypoint spheres* \mathcal{S} for K keypoints $\{\mathbf{p}_i\}_{i=1}^K$ as:

$$\mathcal{S} = \partial \left(\bigcup_{i=1}^K B(\mathbf{p}_i, r) \right)$$

We adopt the SDF of the keypoint spheres to encode keypoints position for our proposed network, with the definition of SDF as:

$$f(\mathbf{p}) : \mathbb{R}^3 \rightarrow s, \quad (1)$$

where $\mathbf{p} \in \mathbb{R}^3$ is an arbitrary point in the space, and $s = \text{sign}(\mathbf{p}) \cdot d$. Here, d represents the distance from \mathbf{p} to the closest point of the sphere surface. We set $\text{sign}(\mathbf{p})$ 1 for the points outside the spheres and -1 for the inside. According to recent work [SMB*20], f should also satisfy the following Eikonal equation:

$$\|\nabla f(\mathbf{p})\| = 1, \quad \forall \mathbf{p} \in \mathbb{R}^3, a.e. \quad (2)$$

$$f(\mathbf{p}) = 0, \quad \forall \mathbf{p} \in \mathcal{S}, \quad (3)$$

where \mathcal{S} is the keypoint spheres. Our key insight is that we can encode keypoints implicitly with an SDF function f . The function is appropriate for an arbitrary number of keypoints and can be well approximated by a neural network function f_θ .

We now bridge the 3D keypoint estimation and various types of inputs of 3D general objects (e.g., images and point clouds) by modeling f_θ conditioned on input from the specified space \mathcal{X} . Given an observation $x \in \mathcal{X}$, the function takes $(\mathbf{p}, x) \in \mathbb{R}^3 \times \mathcal{X}$ to output an SDF value s , which can be formulated as:

$$f_\theta : \mathbb{R}^3 \times \mathcal{X} \rightarrow s. \quad (4)$$

We regard this function as our implicit keypoint network and utilize advanced neural architectures to optimize the parameters θ .

3.1.2. Network Training

Our network adopts the encoder-decoder architecture used in DeepSDF [PFS*19] to learn the implicit field defined w.r.t keypoint spheres. Given the observation of a point cloud or a single-view image, we randomly sample points in the 3D space (i.e., $[-1, 1]^3$) and fetch them into our network to obtain SDF values. As shown in Figure 2, we utilize different encoders for different inputs (PointNet [QSMG17] for point clouds and ResNet [HZRS16] for single-view images), and use a multi-layer perceptron (MLP) as the decoder of SDF. Positional encoding is also applied for each sampled point \mathbf{p} before concatenating it with the observation features, to help the network learn high-frequency components of the input position. Following the work of NeRF [MST*20], the positional encoding function we use is:

$$\psi(\mathbf{p}) = (\sin(2^0 \pi \mathbf{p}), \cos(2^0 \pi \mathbf{p}), \dots, \sin(2^N \pi \mathbf{p}), \cos(2^N \pi \mathbf{p})). \quad (5)$$

The function ψ is applied separately to each coordinate value of \mathbf{p} . In our experiments, we set $N = 6$.

As mentioned in SIREN [SMB*20], SDF learning benefits from

high-frequency features. Therefore, we adopt a sinusoidal activation function, i.e., $\sigma(\cdot) = \sin(\omega \cdot \cdot)$, where $\omega = 30$ is a specified constant in [SMB*20]. We also follow SIREN to initialize the weights of our decoder MLP.

In our experiment, we adopt SDF loss, gradient loss, and normal loss to supervise the training. Specifically, given an observation x and data pairs $\{(\mathbf{p}_i, s_i)\}$ of the queried points and the corresponding SDF values, the SDF loss is defined as the L_1 difference between the predicted values and the ground truth:

$$L_{SDF} = \sum_{x,i} |f_{\theta}(\mathbf{p}_i, x) - s_i|. \quad (6)$$

According to Eq. (2), the norm of $\nabla_{\mathbf{p}_i} f_{\theta}(\mathbf{p}_i, x)$ should be restricted to 1 over \mathbb{R}^3 . For a point on the sphere surface \mathcal{S} , $\nabla_{\mathbf{p}_i} f_{\theta}(\mathbf{p}_i, x)$ equals to the normal vector \mathbf{n}_i at \mathbf{p}_i , if \mathbf{n}_i can be defined here. Thus we add two losses associated with the gradient of our network:

$$L_{grad} = \sum_{x,i} | \|\nabla_{\mathbf{p}_i} f_{\theta}(\mathbf{p}_i, x)\| - 1 |, \quad (7)$$

$$L_{normal} = \sum_{x, \mathbf{p}_i \in \mathcal{S}} \left(1 - \frac{\nabla_{\mathbf{p}_i} f_{\theta}(\mathbf{p}_i, x) \cdot \mathbf{n}_i}{\|\nabla_{\mathbf{p}_i} f_{\theta}(\mathbf{p}_i, x)\|} \right). \quad (8)$$

L_{normal} is used to penalize the cosine similarity between $\nabla_{\mathbf{p}_i} f_{\theta}(\mathbf{p}_i, x)$ and \mathbf{n}_i .

Our final objective of the training is a weighted sum of the three terms:

$$L = \lambda_1 L_{SDF} + \lambda_2 L_{grad} + \lambda_3 L_{normal}, \quad (9)$$

where $\lambda_1, \lambda_2, \lambda_3$ are the corresponding weights. More details are given in the experimental setup (Section 4.1). Note that the positional encoding, sinusoidal activation, and losses associated with the gradient are combined to use for improving the smoothness of keypoint spheres, which can improve the accuracy of the subsequent keypoint estimation from the inferred SDF.

3.2. Keypoint Extraction

To obtain 3D keypoints from the learned SDF, we utilize the Marching Cubes (MC) algorithm [LC87] to first extract the keypoint spheres mesh from the inferred SDF. However, these spheres might intersect each other, hindering the subsequent keypoint extraction. Notice that sphere detection from a point cloud is a well-studied problem, where Hough transform can be used to acquire spheres from point clouds efficiently and robustly [CVC14]. Inspired by this idea, we take the vertices of intersected spheres as input and utilize a Hough transform-based method to extract the distinct spheres as well as their centers. Note that the keypoint extraction task can be simplified and performed on every mesh-connected component of the output mesh, and the underlying spheres should possess the same radius, which greatly reduces the complexity of the calculation. Therefore, we propose Algorithm 1 to extract the keypoints of an unknown number.

In Step 1 of Algorithm 1, we follow the standard Hough transform to voxelize the bounding box of the input point cloud (i.e. one connected component) with a grid size d and perform sphere center voting for all input points. In Step 2, we find out possible clusters containing sphere centers by clustering bins, whose votes

are beyond a given threshold N_{vote} . Then we select points with the maximum votes in each cluster as candidate sphere centers.

In our experiment, the candidate centers calculated in Step 2 are sometimes inaccurate. Therefore, in Step 3, we utilize the nearest points to update the positions of these centers at most N_{max} times. Specifically, we formulate it as a best sphere matching problem (Eq. (10)) with the minimum variance for a given point, which has an analytical solution [CRT*17] given by Eq. (11):

$$\min_{c_L} \sum_{i=1}^{N_L} \left(\|x_i^L - c_L\|_2^2 - \frac{1}{N_L} \sum_{j=1}^{N_L} \|x_j^L - c_L\|_2^2 \right)^2, \quad (10)$$

$$c_L = \bar{X}_L + \frac{1}{2} Cov(X_L)^{-1} \gamma. \quad (11)$$

$$\bar{X}_L = \frac{1}{N_L} \sum_{i=1}^{N_L} x_i^L, \quad Cov(X_L) = \frac{1}{N_L} \sum_{i=1}^{N_L} (x_i^L - \bar{X}_L)(x_i^L - \bar{X}_L)^T,$$

$$\gamma = \frac{1}{N_L} \sum_{i=1}^{N_L} (x_i^L - \bar{X}_L)(x_i^L - \bar{X}_L)^T (x_i^L - \bar{X}_L).$$

where $X_L = \{x_i^L\}_{i=1}^{N_L}$ is the given point cloud and c_L is the optimal center with the minimal variance defined in Eq. (10).

It should be noted that in Step 3, there is a possibility of erroneously grouping some close points into different spheres, such as extracting two center points from an ellipsoid-like shape (a non-standard sphere type in our experiments). To rectify this, in Step 4, we merge these close points by calculating their mean position and then return to Step 3 to ensure accurate keypoint extraction.

For the implementation of keypoint estimation, we set the hyper-parameters $d = 1/32$, radius = 0.08, $\epsilon = 0.01$, $N_{vote} = 80$, $N_{max} = 10$. These hyper-parameters remain constant throughout all experiments, which achieves stable performance.

3.3. Implicit Semantic Learning

In the previous section, we proposed a method to predict the location of keypoints. Practically, the semantic labels (e.g., the head of an airplane, the foot of a chair) of keypoints also play a very important role, such as building correspondence across diverse shapes in the same category. Taking the predicted keypoints of the above framework as input, the semantic keypoint labeling can be treated as a per-point classification task.

For each keypoint, a straightforward way is to extract features conditioned on its coordinates and neighborhood and to feed into an MLP-based classifier. Since the input is a complete point cloud, this task is similar to conducting point-wise classification for picking up the semantic keypoints, such as the method RSNet [HWN18] used in [YLL*20b]. However, it is still very challenging when the input is a partial point cloud or a single-view image, due to the difficulty to provide sufficient features for some unobservable keypoints.

Encouraged by the success of implicit representation learning for

Algorithm 1: Keypoint Extraction

Input: $P = \{p_i\}_{i=1}^{N_P}$
Output: $\{c_i\}_{i=1}^K$

Step 1.
centers of bins $\{b_i\}_{i=1}^{N_B}$, length $d \leftarrow$ Voxelize bounding box of P ,
set $\text{vote}(b_i) = 0, \forall i$.
for p_i in P **do**
 $\text{vote}(b_j) += 1, \forall b_j, \|b_j - p_i\|_2 \leq d/2$

Step 2.
Clusters $\{B_i\}_{i=1}^{N_B} \leftarrow$ Clustering $\{b_j | \text{vote}(b_j) > N_{\text{vote}}\}$
 $c_i^{(0)} \leftarrow \arg \max_{b_j} \{\text{vote}(b_j), b_j \in B_i\}, i = 1, 2, \dots, N_B$

Step 3.
for $k = 0$ to N_{max} **do**
 $P_i^{(k)} \leftarrow \{p_j \in P | \|p_j - c_i^{(k)}\|_2 \leq \|p_j - c_l^{(k)}\|_2, \forall l \neq i\}, \forall i = 1, 2, \dots, N_B$
 $c_i^{(k+1)} \leftarrow \text{Eq}(11) |_{X_l = P_i^{(k)}}, \forall i = 1, 2, \dots, N_B$
 if $\max_i \|c_i^{(k+1)} - c_i^{(k)}\|_2 < \epsilon$ **then**
 break

Step 4.
if exists c_i, c_j such that $\|c_i - c_j\|_2 < \text{radius}$ **then**
 Merge all such c_i, c_j with average position, back to **Step 3.**
else
 return $\{c_i\}_{i=1}^K$

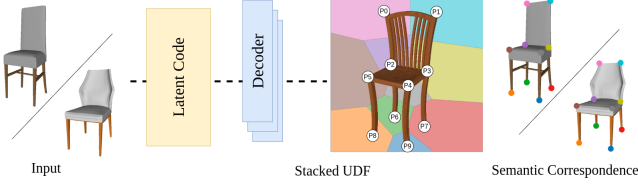


Figure 3: Illustration of the stacked UDF for semantic label prediction. In this case, $\forall \mathbf{p} \in \mathbb{R}^3$, the stacked UDF of \mathbf{p} is $[d_i]_{i=0}^9$, $d_i = \|\mathbf{p} - P_i\|_2, i = 0, \dots, 9$, where P_i is a specific keypoint. Then $\text{Label}(\mathbf{p}) = \arg \min_i [d_i]_{i=0}^9$. We use a 2D Voronoi diagram to illustrate our learned stacked UDF, where the painted region represents points with the same semantic label.

keypoint estimation, we also explore an implicit way to do semantic keypoint learning. Inspired by the stacked fashion of heatmap-based representation [NYD16], we propose our stacked UDF representation. Specifically, assuming the number of semantic labels is K , instead of placing K keypoint labels at one channel of space, we distribute them to K channels of spaces, with one label per channel. For each channel, the distance values from queried points to the specific keypoint of the channel can be computed, as the UDF values.

Stacked UDF learning. We use a neural network to fit stacked UDF for implicit semantic label learning, as shown in Figure 3. Specifically, for an arbitrary point \mathbf{p} in the space, the continuous

stacked UDF g can be defined as:

$$g(\mathbf{p}) : \mathbb{R}^3 \rightarrow d^K \in \mathbb{R}^K, \quad (12)$$

where K is the maximum number of keypoint labels, d^K represents the distance from \mathbf{p} to the corresponding keypoint of every channel. Note that the distance value of the channel should be infinite when the semantic label does not exist. In our experiments, we simply set 1 as the value of the nonexistent label. Given an observation $x \in \mathcal{X}$, we have the conditional network function:

$$g_\theta : \mathbb{R}^3 \times \mathcal{X} \rightarrow d^K \in \mathbb{R}^K. \quad (13)$$

The stacked UDF network shares the same structure as the SDF branch except for the last linear layer in the MLP decoder. This UDF branch is trained using the supervision of L_1 loss on UDF values.

Semantic label prediction. Given the keypoints estimated in Section 3.1, we can obtain their corresponding d^K with learned stacked UDF. Labels of keypoints can then be obtained by using an ‘argmin’ operation to pick out the channel which owns the minimum distance. Notice our method is invariant to the order of input keypoints, making it suitable for unordered keypoint input.

In our semantic learning, we utilize UDF instead of SDF for several reasons. Although using UDF may have a slightly larger fitting error for keypoints compared to using SDF, it still yields comparable results in the task of semantic learning. This is because the accuracy of label prediction is not highly sensitive to the fitting error. Additionally, UDF offers advantages in terms of implementation simplicity, data preparation, and faster inference speed.

Note that in our approach, we separate the tasks of keypoint estimation and semantic learning to achieve optimal performance. An alternative solution is jointly learning keypoint spheres and semantic labels using a similar stacked SDF representation. In this representation, each channel corresponds to the SDF of keypoints associated with a specific label. However, training a stacked SDF model learning requires much more data sampling to ensure accurate keypoint estimation for all channels, significantly increasing the consumption of calculation and storage. Joint learning also may bring more artifacts thus possibly increasing the difficulty of training and the subsequent keypoint extraction. Therefore, we choose to divide the problem into two subtasks to ensure optimal performance.

4. Experiment Results

In this section, we introduce the dataset and implementation details in our experiments (Section 4.1) and evaluate our implicit keypoint learning scheme on the task of keypoint detection (Section 4.2), keypoint estimation (Section 4.3), and semantic label inference (Section 4.4). An ablation study is also conducted for analyzing our architectural design (Section 4.5).

4.1. Experimental Setup

Dataset. We use the KeypointNet dataset [YLL*20b] which contains 103,450 annotated keypoints and 8,234 3D models spanning 16 object categories from ShapeNet [CFG*15]. We choose 10 popular categories, i.e., the airplane, bathtub, car, chair, guitar, knife,

Method	Metric	Airplane	Bath	Chair	Car	Guitar	Knife	Laptop	Motor	Table	Vessel
PointNet	BHD	0.366	0.422	0.310	0.474	0.612	0.890	1.022	0.542	0.660	0.360
	CD	0.070	0.081	0.097	0.198	0.249	0.376	0.552	0.276	0.253	0.110
DGCNN	BHD	0.321	0.354	0.421	0.247	0.183	0.775	0.635	0.474	0.159	0.320
	CD	0.098	0.156	0.119	0.023	0.019	0.113	0.433	0.166	0.073	0.054
Ours	BHD	0.124	0.235	0.148	0.121	0.097	0.147	0.097	0.194	0.117	0.260
	CD	0.015	0.031	0.018	0.009	0.007	0.025	0.015	0.017	0.026	0.053

Table 1: Comparison results of PointNet [QSMG17], DGCNN [WSL*19], and Ours using complete point cloud input. Average BHD and CD are reported, the lower value is better. Our method utilizes the same encoder as PointNet while adopting different keypoint representations. DGCNN is the best keypoint saliency benchmark in KeypointNet [YLL*20b].

Method	Metric	Airplane	Bath	Chair	Car	Guitar	Knife	Laptop	Motor	Table	Vessel
Coords	BHD	0.190	0.249	0.245	0.154	0.126	0.174	0.127	0.219	0.133	0.281
	CD	0.028	0.047	0.060	0.013	0.012	0.036	0.025	0.025	0.025	0.060
Heatmap	BHD	0.241	0.474	0.366	0.163	0.189	0.248	0.834	0.256	0.213	0.611
	CD	0.086	0.345	0.196	0.018	0.030	0.076	1.315	0.049	0.148	0.600
Ours	BHD	0.124	0.235	0.148	0.121	0.097	0.147	0.097	0.194	0.117	0.260
	CD	0.015	0.031	0.018	0.009	0.007	0.025	0.015	0.017	0.026	0.053

Table 2: Comparison results of coordinate regression, heatmap inference, and Ours using complete point cloud input. Average BHD and CD are reported, the lower value is better.

laptop, motorcycle, table, and vessel, for our experiments with point cloud input, and pick rendered images of the corresponding categories in the dataset of 3D-R2N2 [CXG*16] for the experiments using image input. We randomly split the data into a train set (80%) and a test set (20%). More specifically, all the 3D models are normalized into a bounding box of $[-1, 1]^3$. To acquire data pairs for training SDF fields of keypoints, we first represent keypoints as spheres of radius 0.08, created in MeshLab with 2,562 vertices. Subsequently, we uniformly sample 100,000 points in $[-1, 1]^3$ and collect the $N_S \times 2,562$ (N_S is the keypoint number of S) sphere surface points, along with their corresponding SDF values, respectively. For data in semantic learning, samples are kept the same with the SDF setting, and the corresponding distances (i.e., UDF values) to all keypoints will be calculated for each sample point.

Network and training. For the network structure, we deploy classical encoders w.r.t diverse kinds of input followed by the same decoder to estimate keypoints. Specifically, we employ ResNet [HZRS16] for single image input and PointNet [QSMG17] for point cloud input. The dimension of the last layer in all encoders is set to 256. We also incorporate the positional encoding, generating a $39d$ (where d means dimension) feature for each queried point in space. This feature is then concatenated with the $256d$ feature encoded from the input. Subsequently, the $295d$ feature is forwarded to the implicit decoder. The decoder is an MLP-based network consisting of 5 fully connected layers with a sine activation between layers. The output channels are 256, 256, 256, 256, and 1, respectively. During the training of the SDF field, we randomly generate 10,000 volume-based samples and 10,000 surface-based ones (if insufficient, we just add those uniform samples) for each shape. The batch size is set to 4. For training details, in our experiments, we set $\omega = 30$ in the sine function and use $\lambda_1 = 1.0, \lambda_2 = 0.1, \lambda_3 = 0.05$ in the weighted training loss function L . Adam optimizer is adopted to train our network with an initial learning rate of $1e-4$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. We train our network with 300 epochs for all tasks on one GTX 2080ti GPU.

Evaluation metrics. The Chamfer Distance (CD) for traditional

point cloud generation work [FSG17] is adopted to evaluate the distance between the predicted keypoints and the ground truth where the number of points might be different. We also utilize Bidirectional Hausdorff Distance (BHD) to measure the similarity between two point sets. Suppose S_1, S_2 are two point sets, BHD (d_H) and CD (d_C) are defined as:

$$d_H(S_1, S_2) = \frac{1}{2} (\max_{p \in S_1} \min_{q \in S_2} \|p - q\|_2 + \max_{q \in S_2} \min_{p \in S_1} \|p - q\|_2) \quad (14)$$

$$d_C(S_1, S_2) = \frac{1}{|S_1|} \sum_{p \in S_1} \min_{q \in S_2} \|p - q\|_2^2 + \frac{1}{|S_2|} \sum_{q \in S_2} \min_{p \in S_1} \|p - q\|_2^2. \quad (15)$$

4.2. Comparisons on Keypoint Detection

Although our method focuses on keypoint estimation of general objects, especially for incomplete input, we first compare our method with state-of-the-art methods in the field of keypoint detection. Table 1 presents the results of our method, PointNet [QSMG17] (using the same encoder as ours), and DGCNN [WSL*19] (the best of keypoint saliency benchmark in KeypointNet). Both of these classification methods have a tendency to predict an excessive number of keypoints, resulting in significant errors in BHD and CD metrics. Our method outperforms them significantly in terms of BHD and CD, showcasing its superiority.

We further report mIoU curves in line with KeypointNet [YLL*20b] in Figure 4. Our method outperforms DGCNN when the distance threshold is bigger than 0.04. It also achieves better results than an unsupervised keypoint generation method, i.e. UKPGAN [YLL*20a]. The lower mIoU of our method at a small threshold is basically caused by the error of off-surface distance.

Method	Metric	Airplane	Bath	Chair	Car	Guitar	Knife	Laptop	Motor	Table	Vessel
Coords	BHD	0.171	0.209	0.196	0.135	0.135	0.144	0.084	0.201	0.105	0.294
	CD	0.025	0.036	0.036	0.010	0.013	0.023	0.010	0.020	0.019	0.068
Heatmap	BHD	0.236	0.359	0.412	0.141	0.155	0.388	0.616	0.315	0.433	0.504
	CD	0.097	0.248	0.279	0.01	0.026	0.206	0.79	0.087	0.456	0.43
Ours	BHD	0.136	0.209	0.186	0.111	0.097	0.149	0.111	0.187	0.124	0.331
	CD	0.020	0.036	0.032	0.007	0.007	0.030	0.051	0.017	0.029	0.123

Table 3: Comparison results of coordinate regression, heatmap inference, and Ours using partial point cloud input. Average BHD and CD are reported, the lower value is better.

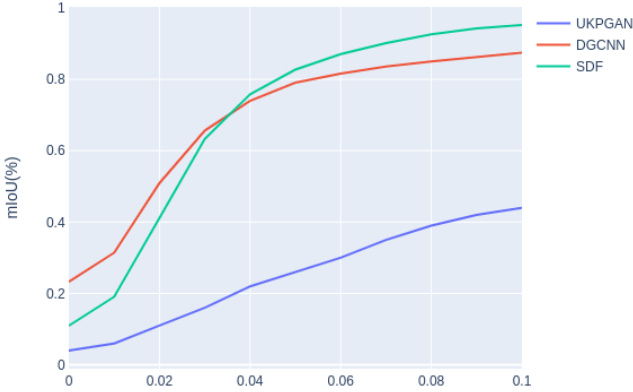


Figure 4: The mIoU results under various distance thresholds (0-0.1) for compared algorithms, i.e., our method (SDF), DGCNN [WSL*19], and UKPGAN [YLL*20a]. Note that our method is significantly better than UKPGAN and outperforms DGCNN when the threshold is larger than 0.04.

4.3. Comparisons on Keypoint Estimation

To validate the effectiveness of our proposed implicit learning approach for 3D keypoint estimation, we conduct experiments on general objects using different input types, i.e., complete point clouds, partial point clouds, and single-view images. The baselines are based on coordinate regression and heatmap inference methods. Because the output dimension of the point regression network is forced to be fixed, we adapt it by learning a binary mask for each predicted keypoint, making it usable for number-varying point prediction. For the heatmap-based method, we represent all keypoints in a 3D heatmap (128^3) to alleviate the calculation and storage consumption. To ensure fairness and efficiency, we employ the same encoder for different methods. Specifically, PointNet [QSMG17] and ResNet18 [HZRS16] are chosen for point cloud and image input, respectively.

Results on complete point clouds. For the complete point cloud input, we employ the same encoder, i.e. PointNet, for different methods. Our method performs the best across almost all categories for all metrics, as demonstrated in Table 2. Visual results are also shown in Figure 5 (the second row).

Results on partial point clouds. We render 24 views of depth maps for each mesh in the KeypointNet dataset [YLL*20b] and then obtain partial point clouds (following the approach used in ME-PCN [GNL*21]) to evaluate the robustness of different methods. The network structures remain unchanged compared to the experiments of complete point clouds. Our method also exhibits its

Method	Metric	Airplane	Chair	Car	Table	Vessel
Coords	BHD	0.247	0.145	0.259	0.140	0.334
	CD	0.052	0.071	0.013	0.035	0.092
Heatmap	BHD	0.389	0.602	0.215	0.652	0.734
	CD	0.181	0.564	0.045	1.079	0.545
Ours	BHD	0.217	0.214	0.136	0.140	0.310
	CD	0.038	0.049	0.012	0.032	0.078

Table 4: Comparison results of coordinate regression, heatmap inference, and Ours using single-view image input. Average BHD and CD are reported, the lower value is better.

superiority in most categories, seen in Table 3 and Figure 5 (the first row). However, when dealing with partial data, both the heatmap method and our approach encounter challenges in accurately predicting the number of keypoints, which can result in potentially larger errors.

Results on single-view images. Similar to single-view reconstruction, the goal of this task is to accurately estimate the complete set of keypoints for a given single image. We evaluate different methods on 5 object categories, in which the rendering images come from 3D-R2N2 [CXG*16]. Quantitative and qualitative results are shown in Table 4 and Figure 6. Our method outperforms others, and the discussion of the partial data effect is similar to the case of the partial point cloud.

4.4. Comparisons on Semantic Learning

In this paper, we introduce an exploring method of semantic label learning for the keypoints generated by our algorithm. Figure 7 illustrates a visual comparison with RSNet [HWN18], which is the best keypoint correspondence benchmark in KeypointNet [YLL*20b]. We evaluate our method using two types of inputs, i.e. full and partial point clouds. To enhance visualization, keypoints with distinct semantic labels are colored in different ways. Our method achieves accurate semantic correspondence with ground truth (GT) for both full and partial point cloud inputs, although there might be some keypoint errors in the case of partial inputs. RSNet demonstrates comparable performance on the full point clouds, but it can have two keypoints with different labels close to each other in the case of partial point cloud inputs. Our method exhibits superiority in handling partial data inputs.

We also present a quantitative comparison between our stacked UDF and RSNet [HWN18] on the settings of complete point cloud input (Table 7) and partial point cloud input (Table 8). As shown in

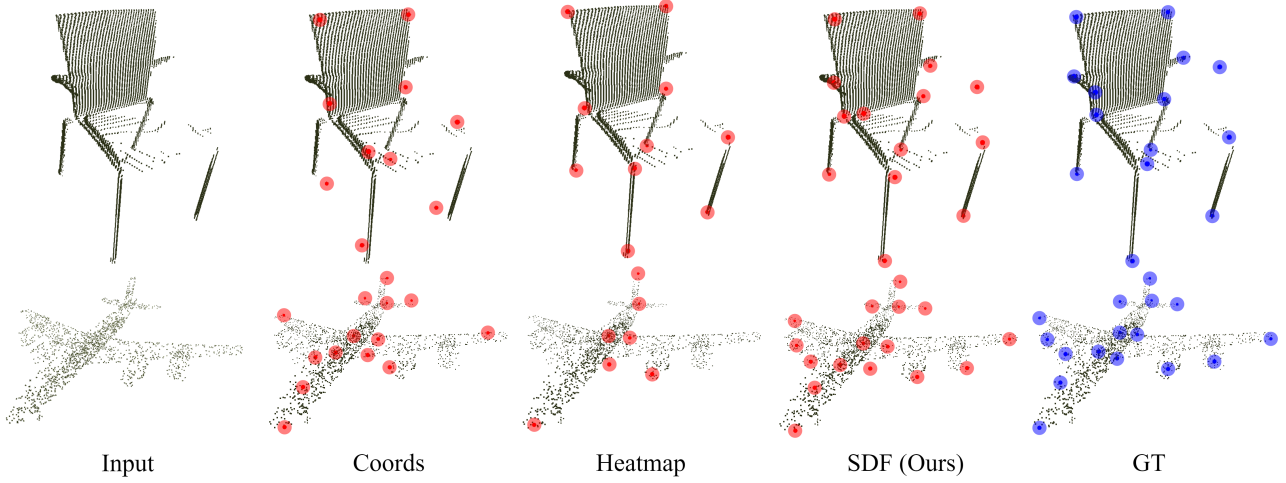


Figure 5: Qualitative results of keypoint estimation with a point cloud input. The input of the first and second rows are a partial point cloud and a complete point cloud, respectively. ‘Coords’ means coordinate regression, and ‘Heatmap’ means heatmap inference.

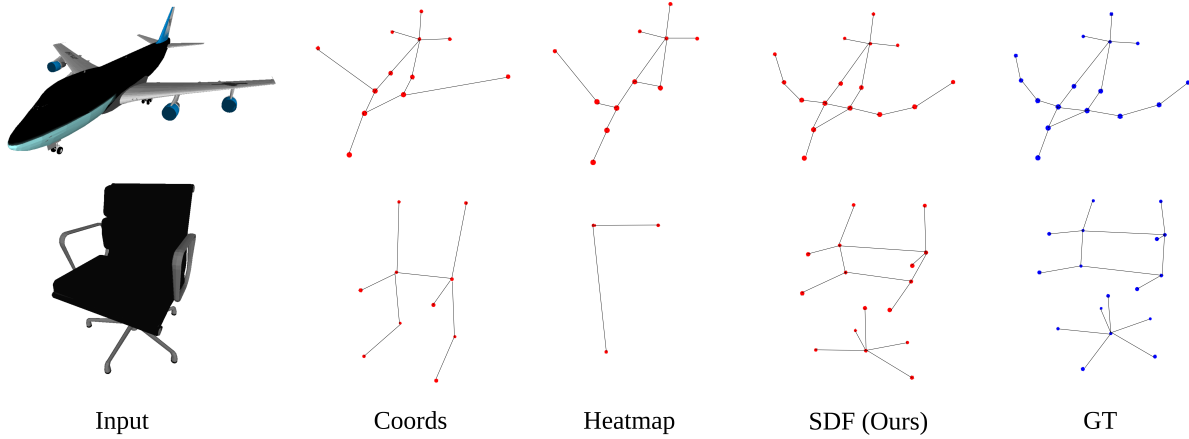


Figure 6: Qualitative results of keypoint estimation with a single-view image input. ‘Coords’ means coordinate regression, and ‘Heatmap’ means heatmap inference. For better visualization, we draw lines among the predicted keypoints.

Metric	Num	BHD	CD
SDF	8	0.0024	1.3e-5
UDF	7	0.0692	0.0119

Table 5: Comparisons of SDF and UDF learned on an example. The ground-truth number of keypoint is 8. The lower of BHD or CD, the result is better.

Radius	0.24	0.16	0.08	0.04	0.02
BHD	0.1507	0.1470	0.1472	0.2148	0.3029
CD	0.0192	0.0181	0.0173	0.0411	0.0696

Table 6: Ablation study of sphere radius. The lower of BHD or CD, the result is better.

these two tables, the Top-1, Top-3, and Top-5 accuracy are reported. RSNet [HWN18] can achieve comparable performance with ours

for complete point cloud input but it degenerates dramatically for partial point cloud input. In contrast, our method can robustly predict semantic labels for both complete and partial data, as stated in Section 3.3 of our paper.

4.5. Ablation Study

In this subsection, we provide an ablation study to validate the effectiveness of our architectural design and hyper-parameter selection. For simplicity without loss of generality, the study is performed on the chair category with complete point cloud input.

Network architecture. We investigate the impact of using positional encoding, gradient loss, and different activation functions in our network training. The results are provided in Table 9, showing the network with positional encoding, gradient loss, and sine activation performs the best. Positional encoding helps the network learn high-frequency features in the data as stated in NeRF [MST*20]

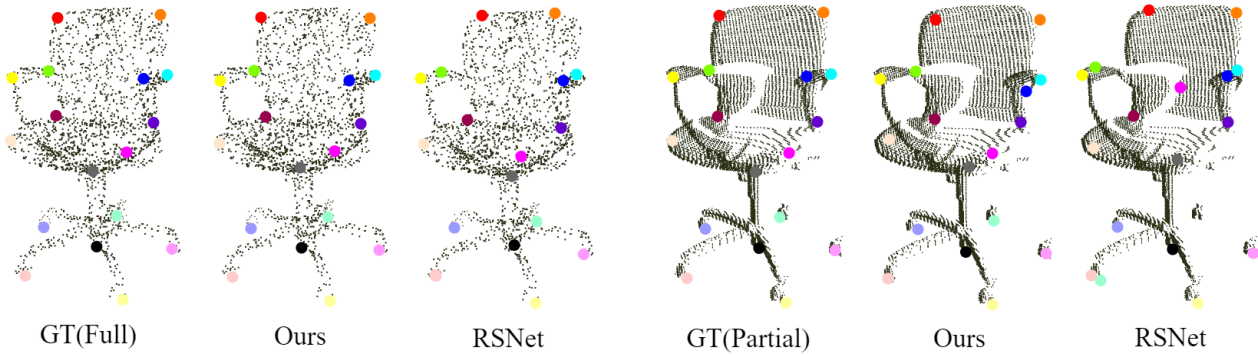


Figure 7: Visualization for results of semantic learning. Left: semantic label prediction with full point cloud input; Right: prediction with the partial point cloud. Our method predicts semantic labels of estimated keypoints by the proposed stacked UDF, while RSNet [HWN18] performs classification on the input point cloud. As seen on the right side, RSNet may generate close keypoints with different semantic labels and keypoints in incorrect positions due to missing data in the input.

Method	Metric	Airplane	Bath	Chair	Car	Guitar	Knife	Laptop	Motor	Table	Vessel	Mean
RSNet	Top-1	49.54	65.14	89.40	60.40	89.78	51.25	96.21	62.77	95.47	80.35	74.03
	Top-3	71.99	87.06	97.97	88.61	98.63	90.00	100.00	81.55	99.39	96.09	91.13
	Top-5	77.90	92.62	99.30	95.49	99.36	100.00	100.00	95.05	99.73	98.41	95.79
Ours	Top-1	76.78	58.17	93.44	80.87	86.09	67.50	100.00	57.16	99.18	49.07	76.82
	Top-3	92.47	82.32	98.40	96.10	98.06	91.67	100.00	79.66	99.62	77.52	91.58
	Top-5	96.02	86.43	98.94	98.74	98.72	98.75	100.00	87.47	99.66	83.23	94.80

Table 7: Comparison results of stacked UDF (Ours) and RSNet [HWN18] with complete point cloud input. The Top-1, Top-3, and Top-5 accuracy are reported. Our method performs slightly better than RSNet.

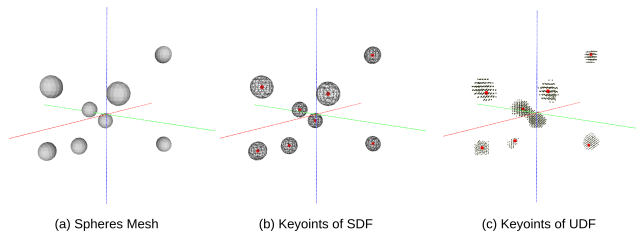


Figure 8: Visualization of fitting results of SDF and UDF. (a) is the spheres' mesh of SDF by Marching Cubes [LC87]. (b) shows the extracted keypoints from the spheres' mesh. (c) shows points whose UDF values are larger than the threshold 0.08 and extracts final keypoints by an 'argmin' function.

and sine activation combined with gradient loss can improve the smoothness of underlying surface discussed in SIREN [SMB*20]. Consequently, these architectural choices contribute to improved keypoint extraction and yield superior results.

SDF vs UDF. As stated in Section 3.1, we adopt SDF for keypoint learning. To compare the regression capabilities of SDF and UDF representations, we use the same MLP network to fit the SDF and UDF fields of 8 points randomly sampled from the $[-1, 1]^3$ space. The numerical results are presented in Table 5. We also give a visualization for the qualitative comparison in Figure 8. For UDF learning, we extract the underlying keypoints from the clusters if their values are larger than the given threshold of 0.08. However, the cluster shapes are not as good as the output of our SDF learn-

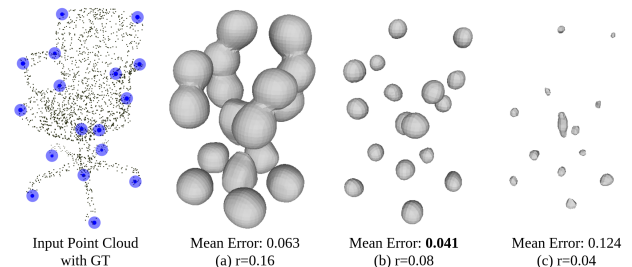


Figure 9: Ablation study about the choice of sphere radius. In this case, mean L_2 distances of keypoints are reported. As shown in the figure, adopting a small radius tends to generate bad shapes. Meanwhile, a large radius leads to generating intersected spheres, increasing the difficulty of keypoint extraction.

ing, making the UDF keypoints extracted from 'argmin' function deviate from the ground truth positions.

Sphere radius. The performance of our SDF representation is influenced by the choice of sphere radius. As shown in Table 6, using a small radius data or a large radius degrades the performance of keypoint estimation. With a small radius, the spheres cannot be well-fitted by the network probably due to local imbalances of SDF values and some numerical issues. On the other hand, using a large radius increases the likelihood of sphere intersections, which complicates keypoint extraction and slows down the process. Figure 9 provides visual results to support these observations. Therefore, we

Method	Metric	Airplane	Bath	Chair	Car	Guitar	Knife	Laptop	Motor	Table	Vessel	Mean
RSNet	Top-1	34.19	40.60	58.66	49.34	69.53	54.79	80.49	48.19	77.96	63.39	57.71
	Top-3	58.37	64.82	78.19	78.28	89.05	88.13	99.43	72.63	97.17	87.84	81.39
	Top-5	67.26	76.44	86.29	88.58	94.29	99.16	100.00	84.47	98.90	94.59	89.00
Ours	Top-1	62.65	49.57	91.86	74.75	84.45	64.38	100.00	56.86	98.07	43.71	72.63
	Top-3	83.29	78.54	97.59	87.87	96.82	91.67	100.00	75.88	99.57	74.18	88.54
	Top-5	88.90	82.50	98.45	93.58	98.19	99.17	100.00	82.45	99.66	82.61	92.55

Table 8: Comparison results of stacked UDF (Ours) and RSNet [HWN18] with partial point cloud input. The Top-1, Top-3, and Top-5 accuracy are reported. It shows that our method is much more robust than RSNet for partial point cloud input.

Activation	ReLU				SeLU				Sine			
	wo,wo	wo,w/	w/,wo	w/,w/	wo,wo	wo,w/	w/,wo	w/,w/	wo,wo	wo,w/	w/,wo	w/,w/
BHD	0.284	0.308	0.227	0.188	0.237	0.207	0.243	0.171	0.216	0.806	0.204	0.148
CD	0.050	0.057	0.038	0.028	0.043	0.036	0.049	0.024	0.031	0.321	0.027	0.018

Table 9: Ablation study of network architecture. We evaluate the effectiveness with (w/) or without (wo) positional encoding (Pos) and gradient loss (Grad), as well as different activation functions (ReLU, SeLU, and Sine). The lower of BHD or CD, the result is better.

adopt a default sphere radius of 0.08, which remains fixed throughout all our experiments.

5. Conclusion

In this paper, we propose a novel framework for general object keypoint estimation, which is the first attempt to introduce continuous implicit field learning into the prediction of sparse and distinct points. It addresses the challenges related to the uncertain number and order properties of keypoints and enhances the performance of 3D keypoint estimation on incomplete input, including partial point clouds and single-view images. Moreover, the proposed implicit representation facilitates semantic label inference. Experimental results demonstrate that our novel keypoint estimation formulation surpasses existing methods that rely on position regression and heatmap inference techniques.

In terms of limitations and future work, our method utilizes a predefined sphere radius for implicit field calculation. It would be valuable to explore the potential benefits of adaptively adjusting the radius based on the specific object categories, which could potentially enhance the accuracy and robustness of keypoint estimation. Additionally, we are intrigued by the prospect of studying dense point scenarios that pose greater challenges for network learning and keypoint extraction. Moreover, an interesting avenue for future research is to develop an end-to-end architecture that directly obtains keypoints from an SDF-based representation, eliminating the need for an intermediate step. Lastly, our method is a supervised approach that relies on an annotated keypoint dataset. In the future, we would like to study unsupervised learning for keypoint estimation which can be generalized to unseen object categories.

Acknowledgment. This work was partially supported by NSFC-62172348, the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, the National Key R&D Program of China with grant No. 2018YFB1800800, by Shenzhen Outstanding Talents Training Fund 202002, by Guangdong Research Projects No.

2017ZT07X152 and No. 2019CX01X104, by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), and by Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. ZDSYS201707251409055). It was also supported in part by Outstanding Young Fund of Guangdong Province with No. 2023B1515020055 and Shenzhen General Project with No. JCYJ20220530143604010. It was also sponsored by CCF-Tencent Open Research Fund.

References

- [BABM19] BARABANAU I., ARTEMOV A., BURNAEV E., MURASHKIN V.: Monocular 3d object detection via geometric reasoning on keypoints. *arXiv preprint arXiv:1905.05618* (2019). 1
- [BADDB11] BERRETTI S., AMOR B. B., DAUDI M., DEL BIMBO A.: 3d facial expression recognition using sift descriptors of automatically detected keypoints. *The Visual Computer* 27, 11 (2011), 1021–1036. 1
- [BF08] BARNEA S., FILIN S.: Keypoint based autonomous registration of terrestrial laser point-clouds. *ISPRS Journal of Photogrammetry and Remote Sensing* 63, 1 (2008), 19–35. 1
- [BKB18] BUGAEV B., KRYSHCHENKO A., BELOV R.: Combining 3d model contour energy and keypoints for object tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 53–69. 1
- [BMSGJL16] BUENO M., MARTÍNEZ-SÁNCHEZ J., GONZÁLEZ-JORGE H., LORENZO H.: Detection of geometric keypoints and its application to point cloud coarse registration. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 49 (2016), 187–194. 1
- [BT16] BULAT A., TZIMIROPOULOS G.: Human pose estimation via convolutional part heatmap regression. In *European Conference on Computer Vision* (2016), Springer, pp. 717–732. 3
- [CAPM20] CHIBANE J., ALLDIECK T., PONS-MOLL G.: Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 6970–6981. 2, 3
- [CC10] CHOI C., CHRISTENSEN H. I.: Real-time 3d model-based tracking using edge and keypoint features for robotic manipulation. In *2010 IEEE International Conference on Robotics and Automation* (2010), IEEE, pp. 4048–4055. 1

- [CFG*15] CHANG A. X., FUNKHOUSER T., GUIBAS L., HANRAHAN P., HUANG Q., LI Z., SAVARESE S., SAVVA M., SONG S., SU H., ET AL.: Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015). 1, 5
- [CPA11] CREUSOT C., PEARS N., AUSTIN J.: Automatic keypoint detection on 3d faces using a dictionary of local shapes. In *2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission* (2011), IEEE, pp. 204–211. 1
- [CRT*17] CERISIER N., REGAD L., TRIKI D., CAMPROUX A.-C., PETITJEAN M.: Cavity versus ligand shape descriptors: Application to urokinase binding pockets. *Journal of Computational Biology* 24, 11 (2017), 1134–1137. 4
- [CVC14] CAMURRI M., VEZZANI R., CUCCHIARA R.: 3d hough transform for sphere recognition on point clouds. *Machine vision and applications* 25, 7 (2014), 1877–1891. 3, 4
- [CXG*16] CHOY C. B., XU D., GWAK J., CHEN K., SAVARESE S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision* (2016), Springer, pp. 628–644. 2, 6, 7
- [CYW*19] CHENG Y., YANG B., WANG B., YAN W., TAN R. T.: Occlusion-aware networks for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 723–732. 2
- [CZ19] CHEN Z., ZHANG H.: Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 5939–5948. 2, 3
- [DGM*19] DABRAL R., GUNDAVARAPU N. B., MITRA R., SHARMA A., RAMAKRISHNAN G., JAIN A.: Multi-person 3d human pose estimation from monocular images. In *2019 International Conference on 3D Vision (3DV)* (2019), IEEE, pp. 405–414. 2
- [DZ19] DOERSCH C., ZISSERMAN A.: Sim2real transfer learning for 3d human pose estimation: motion to the rescue. *Advances in Neural Information Processing Systems* 32 (2019), 12949–12961. 2
- [EMXD19] ESKIMEZ S. E., MADDOX R. K., XU C., DUAN Z.: Noise-resilient training method for face landmark generation from speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2019), 27–38. 2
- [FLCP*20] FERNANDEZ-LABRADOR C., CHHATKULI A., PAUDEL D. P., GUERRERO J. J., DEMONCEAUX C., GOOL L. V.: Unsupervised learning of category-specific symmetric 3d keypoints from point sets. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16* (2020), Springer, pp. 546–563. 2
- [FSG17] FAN H., SU H., GUIBAS L. J.: A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 605–613. 2, 6
- [GNL*21] GONG B., NIE Y., LIN Y., HAN X., YU Y.: Me-pcn: Point completion conditioned on mask emptiness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 12488–12497. 7
- [GSX*21] GENG Z., SUN K., XIAO B., ZHANG Z., WANG J.: Bottom-up human pose estimation via disentangled keypoint regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 14676–14686. 1
- [HSH*20] HE Y., SUN W., HUANG H., LIU J., FAN H., SUN J.: Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 11632–11641. 2
- [HWN18] HUANG Q., WANG W., NEUMANN U.: Recurrent slice networks for 3d segmentation of point clouds. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 2626–2635. 4, 7, 8, 9, 10
- [HZRS16] HE K., ZHANG X., REN S., SUN J.: Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778. 3, 6, 7
- [KKA19] KOCABAS M., KARAGOZ S., AKBAS E.: Self-supervised learning of 3d human pose using multi-view geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 1077–1086. 2
- [KZK17] KHOURY M., ZHOU Q.-Y., KOLTUN V.: Learning compact geometric features. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 153–161. 2
- [LC87] LORENSEN W. E., CLINE H. E.: Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics* 21, 4 (1987), 163–169. 3, 4, 9
- [LHM*15] LI H., HUANG D., MORVAN J.-M., WANG Y., CHEN L.: Towards 3d face recognition in the real: a registration-free approach using fine-grained matching of 3d keypoint descriptors. *International Journal of Computer Vision* 113, 2 (2015), 128–142. 1
- [LL19] LI J., LEE G. H.: Usip: Unsupervised stable interest point detection from 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 361–370. 2
- [LWT20] LIU Z., WU Z., TÓTH R.: Smoke: Single-stage monocular 3d object detection via keypoint estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2020), pp. 996–997. 1
- [MBO06] MIAN A. S., BENNAMOUN M., OWENS R.: Three-dimensional model-based object recognition and segmentation in cluttered scenes. *IEEE transactions on pattern analysis and machine intelligence* 28, 10 (2006), 1584–1601. 1
- [MON*19] MESCHEDER L., OECHSLE M., NIEMEYER M., NOWOZIN S., GEIGER A.: Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 4460–4470. 2, 3
- [MST*20] MILDENHALL B., SRINIVASAN P. P., TANCIK M., BARRON J. T., RAMAMOORTHI R., NG R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision* (2020), Springer, pp. 405–421. 3, 8
- [NN07] NOVATNACK J., NISHINO K.: Scale-dependent 3d geometric features. In *2007 IEEE 11th International Conference on Computer Vision* (2007), IEEE, pp. 1–8. 2
- [NYD16] NEWELL A., YANG K., DENG J.: Stacked hourglass networks for human pose estimation. In *European conference on computer vision* (2016), Springer, pp. 483–499. 5
- [ORL18] OBERWEGER M., RAD M., LEPETIT V.: Making deep heatmaps robust to partial occlusions for 3d object pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 119–134. 2, 3
- [PC15] PFISTER T., CHARLES J., ZISSERMAN A.: Flowing convnets for human pose estimation in videos. In *Proceedings of the IEEE international conference on computer vision* (2015), pp. 1913–1921. 3
- [PFS*19] PARK J. J., FLORENCE P., STRAUB J., NEWCOMBE R., LOVEGROVE S.: DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 165–174. 2, 3
- [PZDD17] PAVLAKOS G., ZHOU X., DERPANIS K. G., DANIILIDIS K.: Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 7025–7034. 3
- [PZK*17] PAPANDREOU G., ZHU T., KANAZAWA N., TOSHEV A., TOMPSON J., BREGLER C., MURPHY K.: Towards accurate multi-person pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 4903–4911. 1

- [QSMG17] QI C. R., SU H., MO K., GUIBAS L. J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 652–660. 3, 6, 7
- [SB11] SIPIRAN I., BUSTOS B.: Harris 3d: a robust extension of the harris operator for interest point detection on 3d meshes. *The Visual Computer* 27, 11 (2011), 963–976. 2
- [SGG*08] SCHALL G., GRABNER H., GRABNER M., WOHLHART P., SCHMALSTIEG D., BISCHOF H.: 3d tracking in unknown environments using on-line keypoint learning for mobile augmented reality. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (2008), IEEE, pp. 1–8. 1
- [SMB*20] SITZMANN V., MARTEL J., BERGMAN A., LINDELL D., WETZSTEIN G.: Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems* 33 (2020). 3, 4, 9
- [SOG09] SUN J., OVSJANIKOV M., GUIBAS L.: A concise and provably informative multi-scale signature based on heat diffusion. In *Computer graphics forum* (2009), vol. 28, Wiley Online Library, pp. 1383–1392. 2
- [SSTN18] SUWAJANAKORN S., SNAVELY N., TOMPSON J., NOROUZI M.: Discovery of latent 3d keypoints via end-to-end geometric reasoning. *arXiv preprint arXiv:1807.03146* (2018). 2
- [SXYL21] SHI R., XUE Z., YOU Y., LU C.: Skeleton merger: an unsupervised aligned keypoint detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 43–52. 2
- [TSDS10] TOMBARI F., SALTI S., DI STEFANO L.: Unique signatures of histograms for local surface description. In *European conference on computer vision* (2010), Springer, pp. 356–369. 2
- [VMB*19] VASCONCELOS L. O., MANCINI M., BOSCAINI D., CAPUTO B., RICCI E.: Structured domain adaptation for 3d keypoint estimation. In *2019 International Conference on 3D Vision (3DV)* (2019), IEEE, pp. 57–66. 2
- [WGY*18] WANG H., GUO J., YAN D.-M., QUAN W., ZHANG X.: Learning 3d keypoint descriptors for non-rigid shape matching. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 3–19. 1
- [WMW*21] WEI G., MA L., WANG C., DESROSIERS C., ZHOU Y.: Multi-task joint learning of 3d keypoint saliency and correspondence estimation. *Computer-Aided Design* 141 (2021), 103105. 2
- [WSL*19] WANG Y., SUN Y., LIU Z., SARMA S. E., BRONSTEIN M. M., SOLOMON J. M.: Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)* 38, 5 (2019), 1–12. 6, 7
- [YLL*20a] YOU Y., LIU W., LI Y.-L., WANG W., LU C.: Ukpgan: Unsupervised keypoint generation. *arXiv preprint arXiv:2011.11974* (2020). 6, 7
- [YLL*20b] YOU Y., LOU Y., LI C., CHENG Z., LI L., MA L., LU C., WANG W.: KeypointNet: A large-scale 3d keypoint dataset aggregated from numerous human annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 13647–13656. 1, 2, 4, 5, 6, 7
- [YWS*21] YUAN Y., WEI S.-E., SIMON T., KITANI K., SARAGIH J.: Simpoe: Simulated character control for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 7159–7169. 2
- [ZHDQ08] ZOU G., HUA J., DONG M., QIN H.: Surface matching with salient keypoints in geodesic scale space. *Computer Animation and Virtual Worlds* 19, 3–4 (2008), 399–410. 1
- [Zho09] ZHONG Y.: Intrinsic shape signatures: A shape descriptor for 3d object recognition. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops* (2009), IEEE, pp. 689–696. 2
- [ZKG*18] ZHOU X., KARPUR A., GAN C., LUO L., HUANG Q.: Unsupervised domain adaptation for 3d keypoint estimation via view consistency. In *Proceedings of the European conference on computer vision (ECCV)* (2018), pp. 137–153. 2