SNAKE: Shape-aware Neural 3D Keypoint Field

Anonymous Author(s)
Affiliation
Address
email

Abstract

Detecting 3D keypoints from point clouds is important for shape reconstruction, while this work investigates the dual question: can shape reconstruction benefit 3D keypoint detection? Existing methods either seek salient features according to statistics of different orders or learn to predict keypoints that are invariant to transformation. Nevertheless, the idea of incorporating shape reconstruction into 3D keypoint detection is under-explored. We argue that this is restricted by former problem formulations. To this end, a novel unsupervised paradigm named SNAKE is proposed, which is short for **s**hape-aware **n**eur**a**l 3D **ke**ypoint field. Similar to recent coordinate-based radiance or distance field, our network takes 3D coordinates as inputs and predicts implicit shape indicators and keypoint saliency simultaneously, thus naturally entangling 3D keypoint detection and shape reconstruction. We achieve superior performance on various public benchmarks, including standalone object datasets ModelNet40, KeypointNet, SMPL meshes and scene-level datasets 3DMatch and Redwood. Intrinsic shape awareness brings several advantages as follows. (1) SNAKE generates 3D keypoints consistent with human semantic annotation, even without such supervision. (2) SNAKE outperforms counterparts in terms of repeatability, especially when the input point clouds are down-sampled. (3) the generated keypoints allow accurate geometric registration, notably in a zero-shot setting. Codes and models will be released.

1 Introduction

2D sparse keypoints play a vital role in both reconstruction [31] and recognition [21], with scale invariant feature transform (SIFT) [18] being arguably the most important pre-Deep Learning (DL) computer vision algorithm. Altough dense alignment using photometric or featuremetric losses is also successful in various domains [2, 35, 8], sparse keypoints are usually preferred due to compactness in storage/computation and robustness to illumination/rotation. Just like their 2D counterparts, 3D keypoints have also drawn a lot of attention from the community in both pre-DL [13, 34] and DL [15, 1, 37] literature, with various applications in reconstruction [42, 40] and recognition[25, 33].

However, detecting 3D keypoints from raw point cloud data is very challenging due to sampling sparsity. No matter how we obtain raw point clouds (e.g., through RGB-D cameras [39], stereo [4], or LIDAR [10]), they are only a discrete representation of the underlying 3D shape. This fact drives us to explore the question of *whether jointly reconstructing underlying 3D shapes helps 3D keypoint detection*. To our knowledge, former methods have seldom visited this idea. Traditional 3D keypoint detection methods are built upon some forms of first-order (e.g., density in intrinsic shape signature [41]) or second-order (e.g., curvature in mesh saliency [14]) statistics, including sophisticated reformulation like heat diffusion [32]. Modern learning-based methods rely upon the idea of consistency under geometric transformations, which can be imposed on either coordinate like USIP [15] or saliency value like D3Feat [1]. The most related method that studies joint reconstruction

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

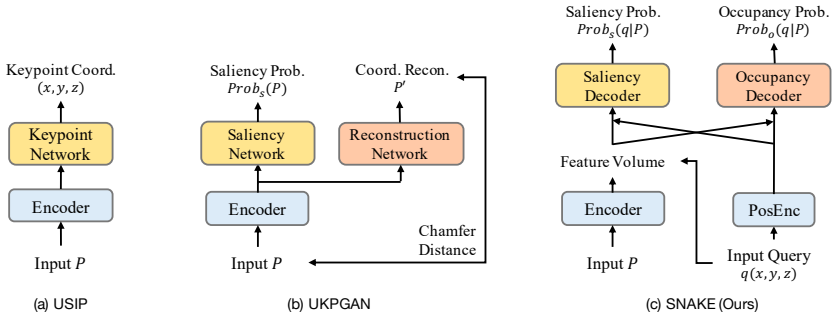

Figure 1: A comparison between existing 3D keypoint detection formulations and our newly proposed one. (a) USIP-like methods directly predict keypoint coordinates from input point clouds P. (b) UKPGAN-like methods predict saliency scores for P. It reconstructs P coordinates simultaneously using chamfer distance. (c) Our SNAKE formulation predicts saliency probabilities and shape indicators for each *continuous* query point q instead of *discrete* point clouds P. Sub-networks used for keypoint detection and reconstruction are shown in yellow and red, although they have different formulations.

and 3D keypoint detection is a recent one named UKPGAN [37], yet it reconstructs input point cloud coordinates using an auxiliary decoder instead of the underlying shape manifold.

Why is this promising idea under-explored in the literature? We argue the reason is that former problem formulations are not naturally applicable for reconstructing the underlying shape surface. Existing paradigms are conceptually illustrated in Fig. 1. USIP-like methods directly output keypoint coordinates while UKPGAN-like methods generate saliency values for input point clouds. In both cases, the representations are based upon *discrete* point clouds. By contrast, we reformulate the problem using coordinate-based networks, as inspired by the recent success of neural radiance fields [20, 16, 28] and neural distance fields [22, 30]. As shown in Fig. 1-c, our model predicts a keypoint saliency value for each *continuous* input query point coordinate $q(x, y, z)$.

A direct advantage of this new paradigm is the possibility of tightly entangling shape reconstruction and 3D keypoint detection. As shown in Fig. 1-c, besides the keypoint saliency decoder, we attach a parallel shape indicator decoder that predicts whether the query point q is occupied. The input to decoders is feature embedding generated by trilinearly sampling representations conditioned on input point clouds P. Imagine a feature embedding at the wing tip of an airplane, if it can be used to reconstruct the sharp curvature of the wing tip, it can be naturally detected as a keypoint with high repeatability. As such, our method is named as **s**hape-aware **n**eur**a**l 3D **ke**ypoint field, or SNAKE.

Shape awareness, as the core feature of our new formulation, brings several advantages. (1) High repeatability. Repeatability is the most important metric for keypoint detection, *i.e.*, an algorithm should detect the same keypoint locations in two-view point clouds. If the feature embedding can successfully reconstruct the same chair junction from two-view point clouds, they are expected to generate similar saliency scores. (2) Robustness to down-sampling. When input point clouds are sparse, UKPGAN-like frameworks can only achieve reconstruction up to the density of inputs. In contrast, our SNAKE formulation can naturally reconstruct the underlying surface up to any resolution because it exploits coordinate-based networks. (3) Semantic consistency. SNAKE reconstructs the shape across instances of the same category, thus naturally encouraging semantic consistency although no semantic annotation is used. For example, intermediate representations need to be similar for successfully reconstructing different human bodies because human shapes are intrinsically similar.

To summarize, this study has the following two contributions:

- We propose a new network for joint shape reconstruction and 3D keypoint detection based upon implicit neural representations. During training, we develop several self-supervised losses that exploit the mutual relationship between two decoders. During testing, we design a gradient-based optimization strategy for maximizing the saliency of keypoints.
- Via extensive quantitative and qualitative evaluations on standalone object datasets ModelNet40, KeypointNet, SMPL meshes, and scene-level datasets 3DMatch and Redwood, we

demonstrate that our shape-aware formulation achieves state-of-the-art performance under three settings: (1) semantic consistency; (2) repeatability; (3) geometric registration.

2 Related Work

3D Keypoint Detector As discussed in the introduction, 3D keypoint detection methods can be mainly categorized into hand-crafted and learning-based. Popular hand-crafted approaches [41, 29, 27] employ local geometric statistics to generate keypoints. These methods usually fail to detect consistent keypoints due to the lack of global context, especially under real-world disturbances, such as density variations and noise. USIP [15] is a pioneering learning-based 3D keypoint detector that outperforms traditional methods by a large margin. However, the detected keypoints are not semantically salient, and the number of keypoints is fixed. Fernandez et al. [9] exploit the symmetry prior to generate semantically consistent keypoints. But this method is category-specific, limiting the generalization to unseen categories and scenes. Recently, UKPGAN [37] makes use of reconstruction to find semantics-aware 3D keypoints. Yet, it recovers explicit coordinates instead of implicit shape indicators. As shown in Fig. 1, different from these explicit keypoint detection methods, we propose a new detection framework using implicit neural fields, which naturally incorporates shape reconstruction.

Implicit Neural Representation Our method exploits implicit neural representations to parameterize a continuous 3D keypoint field, which is inspired by recent studies of neural radiance fields [20, 16, 28] and neural distance fields [22, 30]. Unlike explicit 3D representations such as point clouds, voxels, or meshes, implicit neural functions can decode shapes continuously and learn complex shape topologies. To obtain fine geometry, ConvONet [23] proposes to use volumetric embeddings to get local instead of global features [19] of the input. Recently, similar local geometry preserving networks show a great success for the grasp pose generation [12] and articulated model estimation [11]. They utilize the synergies between their main tasks and 3D reconstruction using shared local representations and implicit functions. Unlike [12, 11] that learn geometry as an auxiliary task, our novel losses tightly couple occupancy and keypoint saliency estimates.

3 Method

This section presents SNAKE, a shape-aware implicit network for 3D keypoint detection. SNAKE conditions two implicit decoders (for shape and keypoint saliency) on shared volumetric feature embeddings, which is shown in Fig. 2-framework. To encourage repeatable, uniformly scattered, and sparse keypoints, we employ several self-supervised loss functions which entangle the predicted occupancy and keypoint saliency, as depicted in the middle panel of Fig. 2. During inference, query points with high saliency are further refined by gradient-based optimization since the implicit keypoint field is continuous and differentiable, which is displayed in Fig. 2-inference.

3.1 Network Architecture

Point Cloud Encoder As fine geometry is essential to local keypoint detection, we adopt the ConvONets [23], which can obtain local details and scale to large scenes, as the point cloud encoder denoted $f_{\theta_{en}}$ for SNAKE. Given an input point cloud $P \in \mathbb{R}^{N\times 3}$, our encoder firstly processes it with the PointNet++ [24] to get a feature embedding $Z \in \mathbb{R}^{N\times C_1}$, where N and C_1 are respectively the number of points and the dimension of the features. Then, these features are projected and aggregated into structured volume $Z' \in \mathbb{R}^{C_1\times H\times W\times D}$, where H, W and D are the number of voxels in three orthogonal axes. The volumetric embeddings serve as input to the 3D UNet [6] to further integrate local and global information, resulting in the output $G \in \mathbb{R}^{C_2\times H\times W\times D}$, where C_2 is the output feature dimension. More details can be found in the supplementary.

Shape Implicit Decoder As shown in the top panel of Fig. 2, each point $q \in \mathbb{R}^3$ from a query set Q is enconded into a C_e-dimensional feature vector q_e via a multi-layer perceptron that is denoted the positional encoder $f_{\theta_{pos}}$, *i.e.* $q_e = f_{\theta_{pos}}(q)$. Then, the local feature G_q is retrieved from the feature volume G according to the coordinate of q via trilinear interpolation. The generated q_e and G_q are concatenated and mapped to the occupancy probability $Prob_o(q|P) \in [0, 1]$ by the occupancy decoder f_{θ_o}, as given in Eq. (1). If q is occupied, the $Prob_o(q|P)$ would be 1, otherwise be 0.

$$f_{\theta_o}(q_e, G_q) \rightarrow Prob_o(q|P) \tag{1}$$

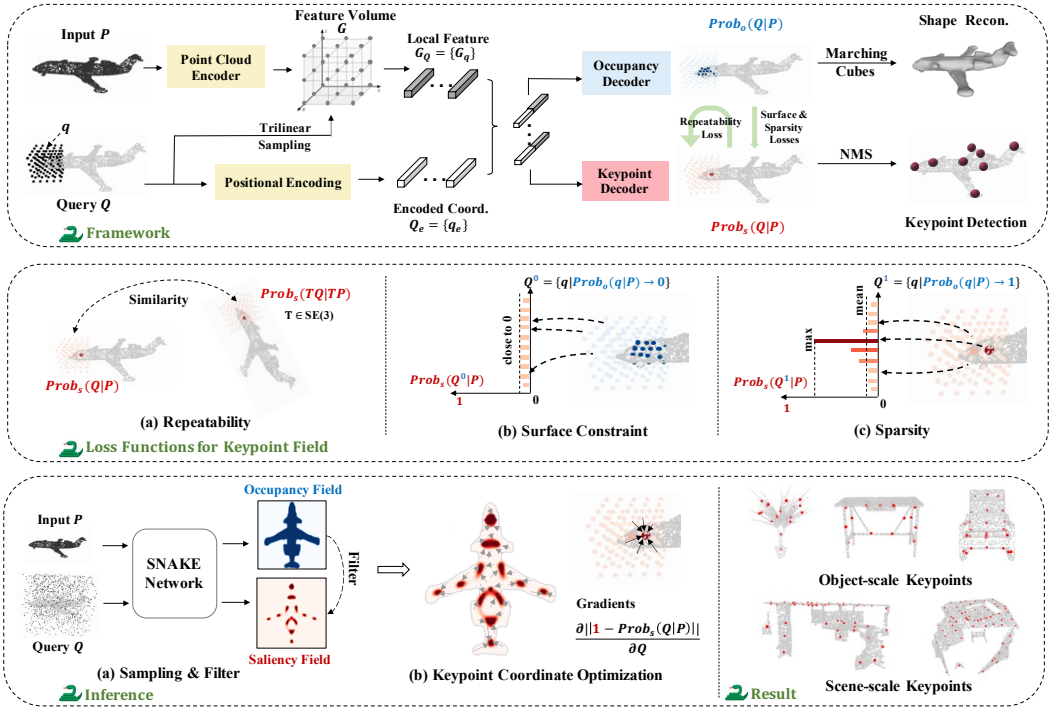

Figure 2: **Framework:** We use an implicit network to decode the occupancy and keypoint saliency probability simultaneously. Green arrows indicate the mutual relationships between the geometry and saliency field. Through marching cubes and non-maximum suppression (NMS), it could respectively recover the shape and detect keypoints from the input. **Loss functions for keypoint filed:** Three loss functions try to make the generated keypoint repeatable, located on the underlying surface, and sparse. **Inference:** We design a gradient-based optimization method to extract keypoints from the saliency field. **Result:** The object-scale and scene-scale keypoints after inference are displayed.

Keypoint Implicit Decoder Most of the process here is the same as in shape implicit decoder, except for the last mapping function. The goal of keypoint implicit decoder f_{θ_s} is to estimate the saliency of the query point q conditioned on input points P, which is denoted as $Prob_s(q|P) \in [0, 1]$ and formulated by:

$$f_{\theta_s}(q_e, G_q) \to Prob_s(q|P). \tag{2}$$

Here, saliency of the query point q is the likelihood that it is a keypoint.

3.2 Implicit Field Training

The implicit field is jointly optimized for occupancy and saliency estimation by several self-supervised losses. In contrast to former arts [12, 11] with a similar architecture that learn multiple tasks separately, we leverage the geometry knowledge from shape field to enhance the performance of keypoint field, as shown in the green arrows of Fig. 2. Specifically, the total loss is given by:

$$\mathcal{L} = \mathcal{L}_o + \mathcal{L}_r + \mathcal{L}_m + \mathcal{L}_s, \tag{3}$$

where $\mathcal{L}_o$ encourages the model to learn the shape from the sparse input, $\mathcal{L}_r$, $\mathcal{L}_m$ and $\mathcal{L}_s$ respectively help the predicted keypoint to be repeatable, located on the underlying surface and sparse.

Occupancy Loss The binary cross-entropy loss l_{BCE} between the predicted occupancy $Prob_o(q|P)$ and the ground-truth label $Prob_o^{gt}$ is used for shape recovery. The queries Q are randomly sampled from the whole volume size $H \times W \times D$. The average over all queries is as follows:

$$\mathcal{L}_o = \frac{1}{|Q|} \sum_{q \in Q} l_{\text{BCE}}\big(Prob_o(q|P), Prob_o^{gt}(q|P)\big), \tag{4}$$

Algorithm 1 Optimization for Explicit Keypoint Extraction

Require: $P, Q_{\text{infer}}, f_{\theta_{en}}, f_{\theta_{pos}}, f_{\theta_o}, f_{\theta_s}$. Hyper-parameters: λ, J, thr_o, thr_s.
  Get initial $Prob_o(Q_{\text{infer}}|P)$ according to Eq.(1).
  Filter to get new query set $Q_{\text{infer}'} = \{q|q \in Q_{\text{infer}}, Prob_o(q|P) > 1 - thr_o\}$.
  for 1 to J **do**
    Evaluate energy function $E(Q_{\text{infer}'}, P)$.
    Update coordinates with gradient descent: $Q_{\text{infer}'} = Q_{\text{infer}'} - \lambda \nabla_{Q_{\text{infer}'}} E(Q_{\text{infer}'}, P)$.
  end for
  Sample final keypoints $Q_k = \{q|q \in Q_{\text{infer}'}, Prob_s(q|P) > thr_s\}$.

where $|Q|$ is the number of queries Q.

Repeatability Loss Detecting keypoints with high repeatability is essential for downstream tasks like registration between two-view point clouds. That indicates the positions of keypoint are covariant to the rigid transformation of the input. To achieve a similar goal, 2D keypoint detection methods [26, 7] enforce the similarity of corresponding local salient patches from multiple views. Inspired by them, we enforce the similarity of local overlapped saliency fields from two-view point clouds. Since the implicit field is continuous, we uniformly sample some values from a local field to represent the local saliency distribution. Specifically, as shown in the top and the middle part of Fig. 2, we build several local 3D Cartesian grids $\{Q_i\}_{i=1}^n$ with resolution of $H_l \times W_l \times D_l$ and size of $1/U$. We empirically set the resolution of Q_i to be almost the same as the feature volume G. As non-occupied regions are uninformative, the center of Q_i is randomly sampled from the input. Then, we perform random rigid transformation T on the P and Q_i to generate TP and TQ_i. Similar to [26], the cosine similarity, denoted as cosim, is exploited for the corresponding saliency grids of Q_i and TQ_i:

$$\mathcal{L}_r = 1 - \frac{1}{n} \sum_{i \in n} \text{cosim}\big(Prob_s(Q_i|P), Prob_s(TQ_i|TP)\big). \tag{5}$$

Surface Constraint Loss As discussed in [15], 3D keypoints are encouraged to close to the input. They propose a loss to constrain the distance between the keypoint and its nearest neighbor from the input. Yet, the generated keypoints are inconsistent when given the same input but with a different density. Thanks to the shape decoder, SNAKE can reconstruct the underlying surface of the input, which is robust to the resolution change. Hence, we use the occupancy probability to represent the inverse distance between the query and the input. As can be seen in Fig. 2-(surface constraint), we enforce the saliency of the query that is far from input P close to 0, which is defined as

$$\mathcal{L}_m = \frac{1}{|Q|} \sum_{q \in Q} \big(1 - Prob_o(q|P)\big) \cdot Prob_s(q|P). \tag{6}$$

Sparsity Loss Similar to 2D keypoint detection methods [26], we design a sparsity loss to avoid the trivial solution ($Prob_s(Q|P)$=0) in Eq.(5)(6). As can be seen in Fig. 2, the goal is to maximize the local peakiness of the local saliency grids. As the sailency values of non-occupied points are enforced to 0 by $\mathcal{L}_m$, we only impose the sparsity loss on the points with high occupancy probability. Hence, we derive the sparsity loss with the help of decoded geometry by

$$\mathcal{L}_s = 1 - \frac{1}{n} \sum_{i \in n} \big(\max Prob_s(Q_i^1|P) - \text{mean}\, Prob_s(Q_i^1|P)\big), \tag{7}$$

where $Q_i^1 = \{q|q \in Q_i, Prob_o(q|P) > 1 - thr_o\}$, $thr_o \in (0, 0.5]$ is a constant, and n is the number of grids. It is noted that the spatial frequency of local peakiness is dependent on the grid size $1/U$, see 4.4. Since the network is not only required to find sparse keypoints, but also expected to recover the object shape, it would generate high saliency at the critical parts of the input, like joint points of a desk and corners of a house, as shown in the Fig. 2-result.

3.3 Explicit Keypoint Extraction

The query point q whose saliency is above a predefined threshold $thr_s \in (0, 1)$ would be selected as a keypoint at the inference stage. Although SNAKE can obtain the saliency of any query point, a higher resolution query set results in a high computational cost. Hence, as shown in Fig. 2-inference,

we build a relatively low-resolution query sets Q_{infer} which are evenly distributed in the input space and further refine the coordinates of Q_{infer} by gradient-based optimization on this energy function:

$$E(Q_{\text{infer}}, P) = \frac{1}{|Q_{\text{infer}}|} \sum_{q \in Q_{\text{infer}}} 1 - Prob_s(q|P). \tag{8}$$

Specifically, details of the explicit keypoint extraction algorithm are summarized in Alg. 1.

4 Experiment

In this section, we evaluate SNAKE under three settings. First, we compare keypoint semantic consistency across **different instances** of the same category, using both rigid and deformable objects. Next, keypoint repeatability of the **same instance** under disturbances such as SE(3) transformation, noise and downsample is evaluated. Finally, we inspect the point cloud registration task on the 3DMatch benchmark, notably in a zero-shot generalization setting. Besides, an ablation study is done to verify the effect of each design choice in SNAKE. The implementation details and hyper-parameters for SNAKE in three settings can be found in the supplementary.

4.1 Semantic Consistency

Datasets The KeypointNet [38] dataset and meshes generated with the SMPL model [17] are utilized. KeypointNet has numerous human-annotated 3D keypoints for 16 object categories from ShapeNet [3]. The training set covers all categories that contain 5500 instances. Following [37], we evaluate 630 unseen instances from airplanes, chairs, and tables. SMPL is a skinned vertex-based deformable model that accurately captures body shape variations in natural human poses. We use the same strategy in [37] to generate both training and testing data.

Metric Mean Intersection over Union (mIoU) is adopted to show whether the keypoints across intra-class instances have the same semantics or not. For KeypointNet, a predicted keypoint is considered the same as a human-annotated semantic point if the geodesic distance between them is under some threshold. Due to the lack of human-labeled keypoints on SMPL, we compare the keypoint consistency in a pair of human models. A keypoint in the first model is regarded semantically consistent if the distance between its corresponding point and the nearest keypoint in the second model is below some threshold.

Evaluation and Results We compare SNAKE with random detection, hand-crafted detectors: ISS [41], Harris-3D [29] and SIFT-3D [27], and DL-based unsupervised detectors: USIP [15] and UKPGAN [37]. As USIP has not performed semantic consistency evaluations, we train the model with the code they provided. We follow the same protocols in [37] to filter the keypoints via NMS with a Euclidean radius of 0.1. Quantitative results are provided in Fig. 5-(a,e). SNAKE obtains higher mIoU than other methods under most thresholds on KeypointNet and SMPL. Qualitative results in Fig. 3 show our keypoints make good alignment with human annotations. Fig. 4 provides qualitative comparisons of semantically consistent keypoints on rigid and deformable objects. Owing to entangling shape reconstruction and keypoint detection, SNAKE can extract aligned representation for intra-class instances. Thus, our keypoints better outline the object shapes and are more semantically consistent under large shape variations. As shown in the saliency field slices, we can get symmetrical keypoints, although without any explicit constraint like the one used in [37].

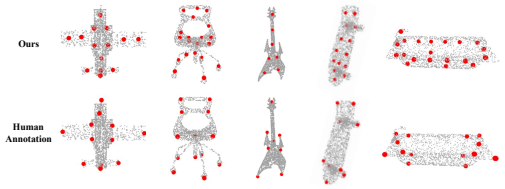

Figure 3: Comparison with human annotations on KeypointNet [38] dataset.

4.2 Repeatability

Datasets ModelNet40 [36] is a synthetic object-level dataset that contains 12,311 pre-aligned shapes from 40 categories, such as plane, guitar, and table. We adopt the official dataset split strategy. 3DMatch [40] and Redwood [5] are RGB-D reconstruction datasets for indoor scenes. Following [15], we train the model on 3DMatch and test it on Redwood to show the generalization performance. The training set contains around 19k samples and the test set consists of 207 point clouds.

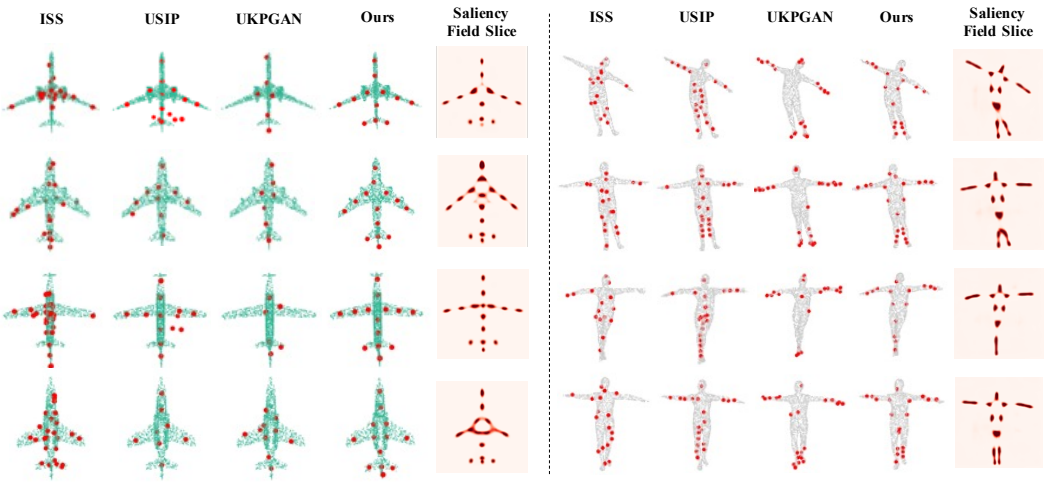

Figure 4: Semantic consistency of keypoints on rigid and deformable objects. Our keypoints are more evenly scattered on the underlying surface of objects, more symmetrical, and more semantically consistent under significant shape variations when compared to other methods. The saliency field slice shows that SNAKE decodes well-aligned saliency values for keypoints in different instances but with similar semantics, such as the wingtip of the airplane and the leg of the human. Here, small saliency is shown in bright red and gets darker with a larger value.

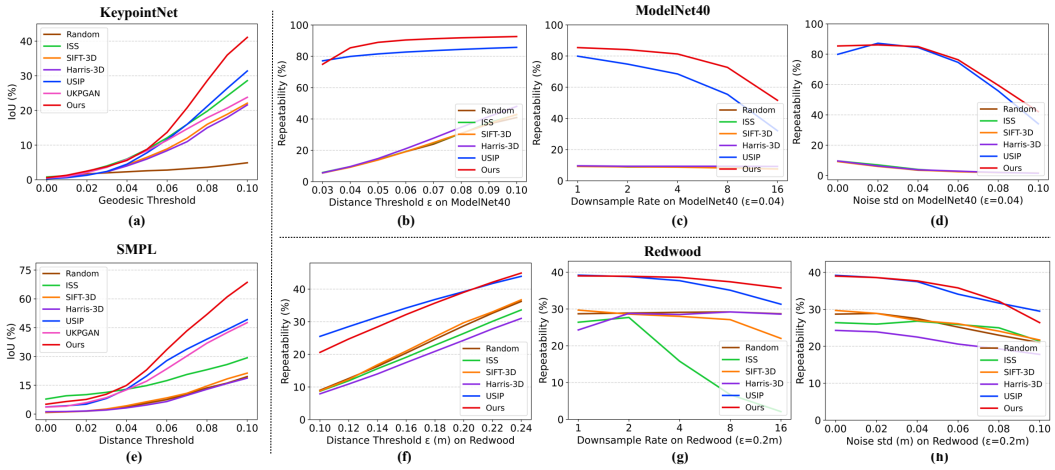

Figure 5: Quantitative results on four datasets. Keypoint semantic consistency (a)(e) on KeypointNet and SMPL. Relative repeatability for two-view point clouds with different distance threshold (b), downsample rate (c), Gaussian noise $\mathcal{N}(0, \sigma_{noise})$ (d) on ModelNet40. The results of (f)(g)(h) are tested on Redwood with the same settings in (b)(c)(d).

Metric We adopt the relative repeatability proposed in USIP [15] as the evaluation metric. Given two point clouds captured from different viewpoints, a keypoint in the first point cloud is repeatable if its distance to the nearest keypoint in the other point cloud is below a threshold ϵ. *Relative* repeatability means the number of repeatable points divided by the total number of detected keypoints.

Evaluation and Results Random detection, traditional methods and USIP are chosen as our baselines. Since UKPGAN does not provide pre-trained models on these two datasets, we do not report its results. We use NMS to select the local peaky keypoints with a small radius (0.01 normalized distance on ModelNet40 and 0.04 meters on Redwood) for ours and baselines. We generate 64 keypoints in each sample and show the performance under different distance thresholds ϵ, downsample rates, and Gaussian noise scales. We set a fixed ϵ of 0.04 normalized distance and 0.2 meters on the ModelNet40 and Redwood dataset when testing under the last two cases. As shown in Fig. 5-(b,f), SNAKE outperforms state-of-the-art at most distance thresholds. We do not surpass USIP on

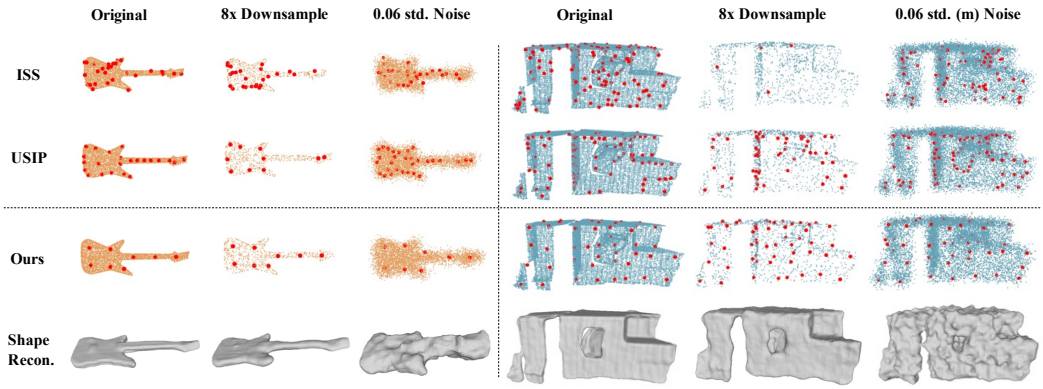

Figure 6: Visualization of keypoints under some disturbances on object-level [36] and scene-level [5] datasets compared to hand-crafted [41] and explicit representation based [15] methods. Downsample rate is 8x and the Gaussian noise scale is 0.06. The shape reconstruction via marching cubes for our occupancy field is also given.

Redwood in the lower thresholds. Note that it is challenging to get higher repeatability on Redwood because the paired inputs have very small overlapping regions. Fig. 5-(c,d,g,h) show the repeatability robustness to different downsample rates and noise levels. SNAKE gets the highest repeatability in most cases because the shape-aware strategy helps the model reason about the underlying shapes of the objects/scenes, which makes keypoints robust to the input variations. Fig. 6 provides visualization of object-level and scene-level keypoints of the original and disturbed inputs. SNAKE can generate more consistent keypoints than other methods under drastic change of inputs.

4.3 Zero-shot Point Cloud Registration

Datasets We follow the same protocols in [37] to train the model on KeypointNet and then directly test it on 3DMatch [40] dataset, evaluating how well two-view point clouds can be registered. The test set consists of 8 scenes which include some partially overlapped point cloud fragments and the ground truth SE(3) transformation matrices.

Metric To evaluate geometric registration, we need both keypoint detectors and descriptors. Thus, we combine an off-the-shelf and state-of-the-art descriptor D3Feat [1] with our and other keypoint detectors. Following [37], we compute three metrics: Feature Matching Recall, Inlier Ratio, and Registration Recall for a pair of point clouds.

Evaluation and Results As baselines, we choose random detection, ISS, SIFT-3D, UKPGAN, and D3Feat. Note that D3Feat is a task-specific learning-based detector trained on the 3DMatch dataset, thus not included in this zero-shot comparison. Ours and UKPGAN are trained on the synthetic object dataset KeypointNet only. The results are reported under different numbers of keypoints (*i.e.*, 2500, 1000, 500, 250, 100). The NMS with a radius of 0.05m is used for D3Feat, UKPGAN, and ours. As shown in Table 1, SNAKE outperforms other methods consistently under three metrics. For registration recall and inlier ratio, we achieve significant gains over UKPGAN and other traditional keypoint methods. Notably, when the keypoints are high in numbers, SNAKE even outperforms D3Feat which has seen the target domain. Local shape primitives like planes, corners, or curves may be shared between objects and scenes, so our shape-aware formulation allows a superior generalization from objects to scenes.

Table 1: Registration result on 3DMatch. We combine the off-the-shelf descriptor D3Feat [1] and different keypoint detectors to perform two-view point cloud registration.

		Feature Matching Recall (%)					Registration Recall (%)					Inlier Ratio (%)				
Detector	Descriptor	2500	1000	500	250	100	2500	1000	500	250	100	2500	1000	500	250	100
D3Feat	D3Feat	95.6	94.5	94.3	93.3	90.6	84.4	84.9	82.5	79.3	67.2	40.6	42.7	44.1	45.0	45.6
Random	D3Feat	95.1	94.5	92.8	90.0	81.2	83.0	80.0	77.0	65.5	38.8	38.6	33.6	28.9	23.6	17.3
ISS	D3Feat	95.2	94.4	93.4	90.1	81.0	83.5	79.2	76.0	64.3	37.2	38.2	33.5	28.8	23.9	17.4
SIFT	D3Feat	94.9	94.0	93.0	91.2	81.3	84.0	79.9	76.1	60.9	38.6	38.4	33.6	28.8	23.3	17.4
UKPGAN	D3Feat	94.7	94.2	93.5	92.6	85.9	82.8	81.4	77.1	69.7	47.4	38.8	35.5	34.0	33.1	27.7
Ours	D3Feat	**95.5**	**95.0**	**94.7**	**92.9**	**89.5**	**85.1**	**83.7**	**81.2**	**74.6**	**50.9**	**41.3**	**39.0**	**37.0**	**33.5**	**30.0**

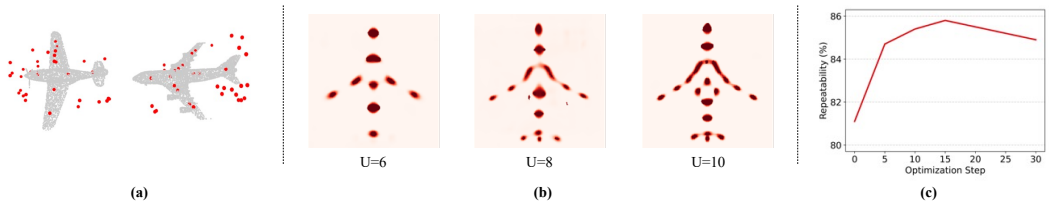

Figure 7: (a) SNAKE fails to predict semantically consistent keypoints without the occupancy decoder. (b) Saliency field slice with a different grid size of $(1/U)^3$. (c) The impact of the optimization step.

4.4 Ablation Study

Loss Function Table 2 reports the performance w.r.t. designs of loss functions. (Row 1) If the occupancy decoder is removed, the surface constraint cannot be performed according to Eq.(6), so they are removed simultaneously. Although the model could detect significantly repeatable keypoints on ModelNet40 [36], it fails to give semantically consistent keypoints on KeypointNet [38]. Fig. 7-a shows that SNAKE is unable to output symmetric and meaningful keypoints without the shape-aware technique. That indicates the repeatability could not be the only criterion for keypoint detection if an implicit formulation is adopted. (Row 2-4) Each loss function for training keypoint field is vital for keypoint detection. Note that the model gives a trivial solution (0) for the saliency field and cannot extract distinctive points when removing the sparsity loss.

Table 2: Ablations for the designs of loss function. occ. = occupancy, sur. = surface, rep. = repeatability, spa. = sparsity and rr. = relative repeatability.

	rr. (%) on [36]			mIoU (%) on [38]		
Threshold ϵ	0.04	0.05	0.06	0.08	0.09	0.1
w/o occ. & sur.	**0.92**	**0.94**	**0.95**	0.22	0.25	0.28
w/o sur.	0.28	0.36	0.42	**0.31**	0.35	0.39
w/o rep.	0.22	0.28	0.34	0.30	0.35	0.39
w/o spa.	0	0	0	0	0	0
w/ all	0.85	0.89	0.90	0.30	**0.37**	**0.42**

Table 3: Impact of different local grid size used in the $\mathcal{L}_o$ and $\mathcal{L}_s$ on ModelNet40.

U	4	6	8	10
rr. (%) (ϵ=0.04)	0.79	**0.85**	0.79	0.77

Table 4: Impact of different global volumetric resolution on ModelNet40.

$H(=W=D)$	32	48	64	80
rr. (%) (ϵ=0.04)	0.62	0.79	**0.85**	0.78

Grid Size and Volumetric Resolution The grid size $1/U$ controls the number of keypoints because $\mathcal{L}_s$ enforces the model to predict a single local maxima per grid of size $(1/U)^3$. Fig. 7-b shows different saliency field slices obtained from the same input with various $1/U$. When U is small, SNAKE outputs fewer salient responses, and more for larger values of U. We also give the relative repeatability results on ModelNet40 under distance threshold $\epsilon = 0.04$ in Table 3, indicating that $U = 6$ gives the best results. From Table 4, we can see that higher resolution improves performance. However, the performance drops when it reaches the resolution of 80. The potential reason is as such: the number of queries in a single grid increases when the resolution becomes higher, as mentioned in 3.2. In this case, finer details make the input to cosine similarity too long and contain spurious values.

Optimization Step and Learning Rate Fig. 7-c shows the importance of optimization (see Alg. 1) for refining keypoint coordinates on the ModelNet40 dataset. It is noted that too many optimization steps will not bring more gains but increase the computational overhead. In this paper, we set the number of update steps to 10. The learning rate for optimization is also key to the final result. When the learning rate is set to 0.1, 0.01, 0.001 and 0.0001, the relative repeatability (%) on ModelNet40 dataset with the same experimental settings as Table 4 are 0.002, 0.622, 0.854 and 0.826, respectively.

5 Conclusion and Discussion

We propose SNAKE, a method for 3D keypoint detection based on implicit neural representations. Extensive evaluations show our keypoints are semantically consistent, repeatable, robust to downsample, and generalizable to unseen scenarios. **Limitations.** The optimization for keypoint extraction during inference requires considerable computational cost and time, which may not be applicable for use in scenarios that require real-time keypoint detection (see supplementary). **Negative Social Impact.** The industry may use the method for pose estimation in autonomous robots. Since our method is not perfect, it may lead to wrong decision making and potential human injury.

References

[1] Xuyang Bai, Zixin Luo, Lei Zhou, Hongbo Fu, Long Quan, and Chiew-Lan Tai. D3feat: Joint learning of dense detection and description of 3d local features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6359–6367, 2020.

[2] Simon Baker and Iain Matthews. Lucas-kanade 20 years on: A unifying framework. *International journal of computer vision*, 56(3):221–255, 2004.

[3] 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.

[4] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. *Advances in Neural Information Processing Systems*, 33:22158–22169, 2020.

[5] Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. Robust reconstruction of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5556–5565, 2015.

[6] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[7] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018.

[8] Jakob Engel, Jurgen Sturm, and Daniel Cremers. Semi-dense visual odometry for a monocular camera. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1456, 2013.

[9] Clara Fernandez-Labrador, Ajad Chhatkuli, Danda Pani Paudel, Jose J Guerrero, Cédric Demonceaux, and Luc Van Gool. Unsupervised learning of category-specific symmetric 3d keypoints from point sets. In *European Conference on Computer Vision*, pages 546–563. Springer, 2020.

[10] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[11] Zhenyu Jiang, Cheng-Chun Hsu, and Yuke Zhu. Ditto: Building digital twins of articulated objects from interaction. *arXiv preprint arXiv:2202.08227*, 2022.

[12] Zhenyu Jiang, Yifeng Zhu, Maxwell Svetlik, Kuan Fang, and Yuke Zhu. Synergies between affordance and geometry: 6-dof grasp detection via implicit representations. *Robotics: science and systems*, 2021.

[13] Andrew E Johnson and Martial Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Transactions on pattern analysis and machine intelligence*, 21(5):433–449, 1999.

[14] Chang Ha Lee, Amitabh Varshney, and David W Jacobs. Mesh saliency. In *ACM SIGGRAPH 2005 Papers*, pages 659–666. 2005.

[15] Jiaxin Li and Gim Hee Lee. Usip: Unsupervised stable interest point detection from 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 361–370, 2019.

[16] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33:15651–15663, 2020.

[17] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6), oct 2015.

[18] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[19] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019.

[20] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020.

[21] David Nister and Henrik Stewenius. Scalable recognition with a vocabulary tree. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2161–2168. Ieee, 2006.

[22] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019.

[23] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *European Conference on Computer Vision*, pages 523–540. Springer, 2020.

[24] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[25] Hossein Rahmani, Arif Mahmood, Q Du Huynh, and Ajmal Mian. Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition. In *European conference on computer vision*, pages 742–757. Springer, 2014.

[26] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. *Advances in Neural Information Processing Systems*, 32, 2019.

[27] Blaine Rister, Mark A Horowitz, and Daniel L Rubin. Volumetric image registration from invariant keypoints. *IEEE Transactions on Image Processing*, 26(10):4900–4910, 2017.

[28] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020.

[29] Ivan Sipiran and Benjamin Bustos. Harris 3d: a robust extension of the harris operator for interest point detection on 3d meshes. *The Visual Computer*, 27(11):963–976, 2011.

[30] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020.

[31] Noah Snavely, Steven M Seitz, and Richard Szeliski. Modeling the world from internet photo collections. *International journal of computer vision*, 80(2):189–210, 2008.

[32] Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Computer graphics forum*, volume 28, pages 1383–1392. Wiley Online Library, 2009.

[33] Supasorn Suwajanakorn, Noah Snavely, Jonathan J Tompson, and Mohammad Norouzi. Discovery of latent 3d keypoints via end-to-end geometric reasoning. *Advances in neural information processing systems*, 31, 2018.

[34] Federico Tombari, Samuele Salti, and Luigi Di Stefano. Performance evaluation of 3d keypoint detectors. *International Journal of Computer Vision*, 102(1):198–220, 2013.

[35] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. Deepflow: Large displacement optical flow with deep matching. In *Proceedings of the IEEE international conference on computer vision*, pages 1385–1392, 2013.

[36] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

[37] Yang You, Wenhai Liu, Yong-Lu Li, Weiming Wang, and Cewu Lu. Ukpgan: Unsupervised keypoint ganeration. *arXiv preprint arXiv:2011.11974*, 2020.

[38] Yang You, Yujing Lou, Chengkun Li, Zhoujun Cheng, Liangwei Li, Lizhuang Ma, Cewu Lu, and Weiming Wang. 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*, pages 13647–13656, 2020.

[39] Aviad Zabatani, Vitaly Surazhsky, Erez Sperling, Sagi Ben Moshe, Ohad Menashe, David H Silver, Zachi Karni, Alexander M Bronstein, Michael M Bronstein, and Ron Kimmel. Intel® realsense™ sr300 coded light depth camera. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2333–2345, 2019.

[40] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas A. Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 199–208, 2017.

[41] Yu Zhong. Intrinsic shape signatures: A shape descriptor for 3d object recognition. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 689–696. IEEE, 2009.

[42] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Fast global registration. In *European conference on computer vision*, pages 766–782. Springer, 2016.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 5.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See supplemental material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See supplemental material.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See supplemental material.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [Yes] See supplemental material.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See supplemental material.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See supplemental material.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See supplemental material.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]