UKPGAN: A General Self-Supervised Keypoint Detector

Yang You, Wenhai Liu, Yanjie Ze, Yong-Lu Li, Weiming Wang*, Cewu Lu*
Shanghai Jiao Tong University, China
{qq456cvb, sjtu-wenhai, zeyanjie, yonglu_li, wangweiming, lucewu}@sjtu.edu.cn

Abstract

Keypoint detection is an essential component for the object registration and alignment. In this work, we reckon keypoint detection as information compression, and force the model to distill out important points of an object. Based on this, we propose UKPGAN, a general ***self-supervised*** *3D keypoint detector where keypoints are detected so that they could reconstruct the original object shape. Two modules:* ***GAN-based keypoint sparsity control*** *and* ***salient information distillation*** *modules are proposed to locate those important keypoints. Extensive experiments show that our keypoints align well with human annotated keypoint labels, and can be applied to SMPL human bodies under various non-rigid deformations. Furthermore, our keypoint detector trained on clean object collections generalizes well to real-world scenarios, thus further improves geometric registration when combined with off-the-shelf point descriptors. Repeatability experiments show that our model is stable under both rigid and non-rigid transformations, with local reference frame estimation. Our code is available on https://github.com/qq456cvb/UKPGAN.*

1. Introduction

Recently, 3D object analysis and scene understanding receive more and more attentions. Though plenty of methods [7, 13, 19, 23] on object analysis have been proposed, there is still a lack of capability of processing and understanding objects, especially under an unsupervised setting.

3D keypoints, unlike part annotations, provide a sparse but meaningful representations of an object. They are widely leveraged in many tasks such as object matching, object tracking, shape retrieval and registration [4,22,34]. Keypoint detections have its origin in 2D image processing [12,21,28]. In 3D domain, traditional methods like Harris-3D [29], HKS [30], Salient Points [5], Mesh Saliency [18], ISS [41], Sift-3D [27] and Scale Dependent Corners [24] propose to detect keypoints based on geometric variations. However, these hand-crafted detectors rely heavily on hard-coded parameters and their performance is not comparable to current learning-based methods.

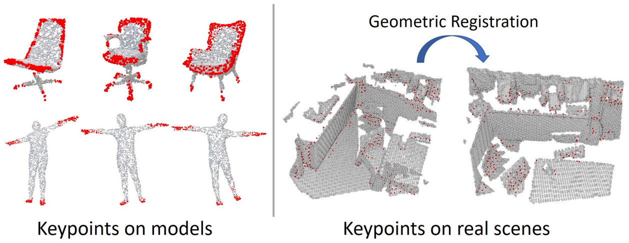

Figure 1. Our model outputs unsupervised keypoints and embeddings given a point cloud, in either rigid or non-rigid deformations. Left are keypoint predictions on clean models (indicated by red) and right are keypoint predictions on real scenes, best viewed in color. These keypoints are *consistent* and could be used for registration.

Recently, some learning-based methods like USIP [20] and D3Feat [2] have been proposed. USIP regresses keypoint locations from pre-segmented local groups and then utilizes a probabilistic chamfer loss. However, their method requires the farthest point sampling and may output points that are not on the input. D3Feat instead gives saliency scores and descriptors densely for each point. Both USIP and D3Feat predict 3D keypoints by solving the auxiliary task of correctly estimating rotations in a Siamese architecture. They both require real-world point clouds during training, and do not have much control on the output keypoints.

To solve these problems, we follow a totally different route to obtain 3D keypoints, which is named **Unsupervised Key Point GANeration (UKPGAN)**. A keypoint saliency distribution is given through a detector network, with a novel adversarial **GAN loss to control its sparsity**. Then, to make these keypoints informative, we leverage a **salient information distillation** process to reconstruct the original point cloud from these sparse keypoints, forming an encoder-decoder architecture. Our model can be seen as an *information compression scheme*, keeping most information of

[*] Cewu Lu and Weiming Wang are the corresponding authors. Cewu Lu is member of Qing Yuan Research Institute and MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China and Shanghai Qi Zhi institute.

the object with the least keypoints. The rationale behind our method is simple but powerful: one should be able to fully recover an object's structure from a small set of keypoints. This also coincides with that mentioned in [35]: "*much of human learning, perception, and cognition, may be understood as information compression*". Results show that our model could output stable informative keypoints from unseen objects, and generalize well to real-world scenarios.

Compared to previous methods, UKPGAN has the following advantages: 1) our detector is proven to be rotation invariant without any data augmentations, by first estimating a Local Reference Frame (LRF), which also makes our local keypoint representation disentangled from rotations; 2) detected keypoints are intra-class consistent and stable on both rigid and non-rigid objects, with high repeatability; 3) our model trained on clean object collections (i.e. ModelNet) generalizes well to real-world point clouds, free from the usage of real-world training data.

We first evaluate our method on ShapeNet models with keypoint labels. Our model achieves remarkable results in keep consistent with human labeled part and keypoints. UKPGAN cannot only be applied to rigid but non-rigid objects by keeping consistency on SMPL human body deformable meshes. As an application of our model, we also evaluate UKPGAN on 3DMatch and ETH datasets, which are real-world geometric registration benchmarks. Experiments show that when trained on clean objects (i.e. ModelNet), our model generalizes well to real-world scenarios, and further improves the registration performance of current state-of-the-art methods. At last, extensive experiments are conducted to demonstrate that UKPGAN achieves high rotation repeatability, which is an important and desired property of keypoints.

2. Related Work

2.1. Hand-crafted Keypoint Detectors

Prior to deep learning, researchers proposed numerous methods to detect stable interest points on objects, in both 2D and 3D domains. SIFT [21], ORB [28] and SURF [3] extract features by detecting local pattern variations on 2D images. They are robust to scale and rotation changes and give consistent keypoints on two identity objects. 3D Harris [29] extends Harris corner detector to 3D meshes. HKS [30] proposes a novel point signature based on the properties of the heat diffusion process on a shape. Salient Points [5] model interest points by a Hidden Markov Model (HMM), which is trained in an unsupervised way by using contextual 3D neighborhood information. Mesh Saliency [18] defines mesh saliency in a scale-dependent manner using a center-surround operator on Gaussian-weighted mean curvatures. CGF [16] learns to represent the local geometry around a point in an unstructured point cloud. 3D SIFT [27] is an analogue of the scale-invariant feature transform (SIFT) for three-dimensional images. ISS [41] introduces intrinsic shape signature, which uses a view-independent representation of the 3D shape to match shape patches from different views directly. However, these methods only consider the local geometric information without semantic knowledge, leading to a discrepancy from human perceptions.

2.2. Learning-based Keypoint Detectors

Recently, some deep learning based detectors have been proposed to bypass hand-crafted keypoint detection rules, in both 2D and 3D domains. On 2D images, some unsupervised keypoint detection methods are proposed. Jakab et al. [14] extracts semantically meaningful keypoints by passing a target image through a tight bottleneck to distill the geometry of the object. Zhang et al. [40] uses an auto-encoding module with channel-wise softmax operation to discover landmarks. Suwajanakorn et al. [32] discover latent 3D keypoints from 2D images by enforcing multi-view consistency. Georgakis et al. [9] employ a Siamese architecture augmented by a sampling layer and a novel score loss function to detect keypoints on depth maps. In 3D domain, methods like SyncSpecCNN [36] and deep functional dictionaries [31] rely on ground-truth keypoint supervision. For unsupervised methods, USIP [20] regresses keypoint locations from pre-segmented local groups and then utilizes a probabilistic Chamfer loss. D3Feat [2] instead gives saliency scores and descriptors densely for each point. It relies on an auxiliary task of correctly estimating rotations in a Siamese architecture, ignoring semantic information. There exists another line of search [6,8,15] that outputs a predefined fixed number of keypoints by regressing the absolute coordinates. However, these methods are not robust to rigid transformations and fail to generalize to real-world scenarios.

3. Approach

3.1. Overview

Given a point set $\mathbf{X} = \{\mathbf{x}_n | \mathbf{x}_n \in \mathbb{R}^3, n = 1, 2, \dots N\}$ with $\mathbf{x}$ sampled from some manifold $\mathcal{M}$, we seek a keypoint set $\tilde{\mathbf{X}} \subseteq \mathbf{X}$, where $|\tilde{\mathbf{X}}|$ is the number of required keypoints.

Here, we propose an unsupervised encoder-decoder architecture. In the encoder, which is also the detector, a keypoint probability s is predicted for each point. To keep the detected keypoints sparse, GAN-based keypoint sparsity control is leveraged. In the decoder, also a reconstruction network, we utilize salient information distillation to reconstruct the original point cloud, in an unsupervised way. The intuition is that, a set of good keypoints should contribute to the unique information of an object, making reconstruction possible. The overview of our method is shown in Figure 2.

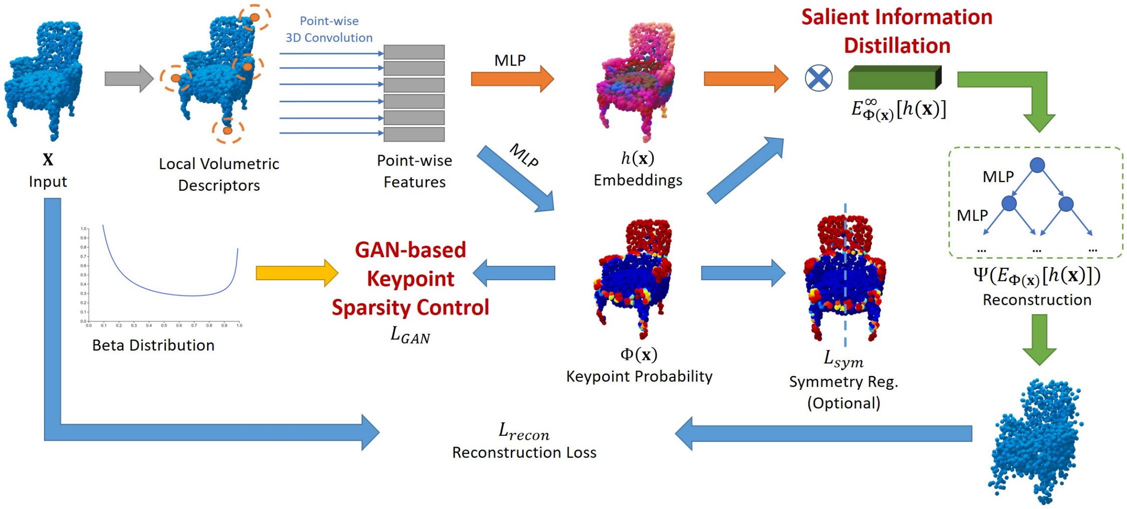

Figure 2. **Our whole pipeline on keypoint and embedding generations.** We first extract point-wise rotation invariant features and then output keypoint probabilities and semantic embeddings through two MLPs, respectively. GAN is leveraged to control keypoint sparsity, and salient information distillation is utilized to distill most salient features. A decoder is concatenated to reconstruct original point clouds.

3.2. Rotation Invariant Feature Extraction

In order to be robust under rigid transformations, we first generate a Local Reference Frame (LRF) by co-variance eigen-decomposition on each point $\mathbf{x}$'s spherical neighborhoods $\mathcal{S} = \{\mathbf{x}_i : \|\mathbf{x}_i - \mathbf{x}\|_2 \leq r\}$. Then points in the local neighborhood $\mathbf{x}_i \in \mathcal{S}$ are transformed to their canonical position $\mathbf{x}'_i$ according to the estimated LRFs. Next, we follow the same strategy as that in PerfectMatch [10] to discretize these points in a Smoothed Density Value (SDV) grid, centered on the point $\mathbf{x}$ and aligned with the LRF. The voxelization is based on Gaussian smoothing kernel. Afterwards, we would get a voxelized descriptor $\mathcal{F}(\mathbf{x}) \in \mathbb{R}^{W\times H\times D}$ for each point $\mathbf{x}$. For mode details, we refer the reader to PerfectMatch [10]. These point-wise 3D descriptors are batched together and fed through 3D convolution layers to be further refined. Thanks to the estimated LRFs, this step provides *local rotation-invariant* features, which are critical for rotation repeatability.

3.3. Dual Branches on Estimating Probabilities and Embeddings

After extracting rotation invariant point-wise features, we use dual Multi-Layer Perceptron (MLP) networks to estimate a keypoint salient probability $\Phi(\mathbf{x}) \in [0, 1]$ and a high-dimensional embedding $h(\mathbf{x}) \in \mathbb{R}^F$, which will be used for reconstruction.

Sparsity on $\Phi(\mathbf{x})$. In order to compress the entire point cloud with a minimum set of keypoints, $\Phi(\mathbf{x})$ needs to be sparse. What is a good way to make $\Phi(\mathbf{x})$ sparse? One would consider L1 regularization. However, it tends to output more probabilities around zero and does not have much control over non-zero probabilities. In order to output distinguishable keypoints and suppress those meaningless points, we would like $\Phi(\mathbf{x})$ to accumulate around both 0's and 1's. A straight-forward solution is to define a controllable keypoint distribution that accumulates around 0's and 1's, then force the network prediction to match this prior. Inspired by [38], we take Beta distribution (shown in Figure 2) as our keypoint distribution prior. In Beta distribution, there are two parameters α and β, which control the accumulation of positive (1) and negative (0) samples, respectively. For more details of controllability provided by Beta distribution, please refer to our supplementary.

GAN-based Keypoint Sparsity Control A direct solution to force the sparsity is to compute the KL divergence between the predicted keypoint distribution and the Beta prior. However, since we are predicting keypoint *samples* instead of distribution *parameters*, the closed form of KL divergence between Beta prior and $\Phi(\mathbf{x})$ does not exist. We resort to adversarial loss to resolve this.

GAN [11] is leveraged to generate fake keypoint distributions that look real to our Beta prior (i.e., $p(\mathbf{x})$). It requires a discriminator network D and a generator network (i.e., $\Phi(\cdot)$). Notice that in our adversarial training settings, each sample is a keypoint distribution on a point cloud, which is itself sampled from a repository. The input to the discriminator network D is the whole keypoint distribution set on a single point cloud. The reason not to input the single keypoint is that we want each object's keypoint distribution to follow our Beta prior, but not the distribution of the points from all objects.

In practice, we employ WGAN-GP [1] instead of the naive GAN loss as it is more robust. The loss follows:

$$L_{GAN} = \min_{\Phi} \max_{D} (\mathbb{E}_{\mathcal{M}}[D(\{p(\mathbf{x})|\mathbf{x} \in \mathcal{M}\})] \tag{1}$$

$$- \mathbb{E}_{\mathcal{M}}[D(\{\Phi(\mathbf{x})|\mathbf{x} \in \mathcal{M}\})] + \lambda(\|\nabla D\|_2 - 1)^2), \tag{2}$$

which penalizes the gradient of the discriminator.

3.4. Reconstruction Network

Given a keypoint distribution $\{\Phi(\mathbf{x}) \in \mathbb{R}|\mathbf{x} \sim \mathcal{M}\}$ and high-dimensional embeddings $\{h(\mathbf{x}) \in \mathbb{R}^F|\mathbf{x} \sim \mathcal{M}\}$, a point cloud decoder is introduced to reconstruct the original shape. Denoting the point cloud decoder as $\Psi : \mathbb{R}^N \times \mathbb{R}^{N\times F} \to \mathbb{R}^{N\times 3}$, the reconstruction loss can be expressed as follows:

$$L_{recon} = CD(\Psi(\{\Phi(\mathbf{x})|\mathbf{x} \sim \mathcal{M}\}, \{h(\mathbf{x})|\mathbf{x} \sim \mathcal{M}\}), \mathbf{X}), \tag{3}$$

where CD is the Chamfer distance.

Salient Information Distillation In Equation 3, Ψ takes both keypoint distribution and high-dimension embeddings as the input. Recall that our goal is to find a sparse set of salient keypoints that possibly reconstruct the original shape. To fulfill this, we get some inspiration from the *max* operation in PointNet [26] and propose a *salient information distillation* module. This module is leveraged to force the network to give both probable (large $\Phi(\mathbf{x})$) and semantic-rich (large $h(\mathbf{x})$) keypoints.

We define Ψ as:

$$\Psi = \text{TopNet}(\max_{\mathbf{x}\sim\mathcal{M}}[\Phi(\mathbf{x}) \cdot h(\mathbf{x})]), \tag{4}$$

where we slightly abuse the notation such that $\Phi(\mathbf{x})$ is broadcasted when multiplying with $h(\mathbf{x})$. The max operation is also conducted channel-wisely, so that $\max_{\mathbf{x}\sim\mathcal{M}}[\Phi(\mathbf{x}) \cdot h(\mathbf{x}) \in \mathbb{R}^F$. TopNet represents the point decoder structure similar to that of Tchapmi *et al.* [33].

In addition, for semantic $h(x)$, we care about the absolute value of $h(\mathbf{x})$ (features with large negative magnitude should not be suppressed) and the final decoder is

$$\Psi = \text{TopNet}(\max_{\mathbf{x}\sim\mathcal{M}}[\Phi(\mathbf{x}) \cdot \max(h(\mathbf{x}), 0)] \tag{5}$$

$$\oplus \max_{\mathbf{x}\sim\mathcal{M}}[\Phi(\mathbf{x}) \cdot \max(-h(\mathbf{x}), 0)]), \tag{6}$$

where $\oplus$ means concatenation.

Intuitively, our decoder forces the network to mark those semantic-rich (large $h(x)$) as salient keypoints (large $\Phi(x)$), otherwise the product will be small and get suppressed because of the *max* operation. On the other side, indistinguishable points with similar local context are therefore discarded. For example, a rectangle can be perfectly reconstructed given the four corners. Points between the corners provide little information about the overall shape. Detailed analysis on salient information distillation will be given in Section 4.5.

3.5. Symmetric Regularization

Although we first extract rotation invariant local descriptors from original point clouds, it is not symmetric invariant. For most common objects, we have a strong prior such that detected keypoints and features should be symmetric, leading to the following loss:

$$L_{sym} = \frac{1}{|\mathbf{S}|} \sum_{(\mathbf{x},\mathbf{x}')\in\mathbf{S}} (\|\Phi(\mathbf{x}) - \Phi(\mathbf{x}')\| + \|h(\mathbf{x}) - h(\mathbf{x}')\|_1), \tag{7}$$

where $\mathbf{S}$ is the set of all symmetric point pairs. Note that symmetric regularization is only used for training; in testing, symmetric information about objects is not required.

The final loss is an empirical sum of three terms:

$$L = \eta_1 \cdot L_{recon} + \eta_2 \cdot L_{GAN} + \eta_3 \cdot L_{sym}. \tag{8}$$

3.6. Implementation Details

Network Architecture Our model takes a point cloud $\mathbf{X} \in \mathbb{R}^{N\times 3}$ as input where $N = 2048$. Then a voxelized descriptor is extracted for each point with $\{\mathcal{F}(\mathbf{x_n})\}_{n=1}^N \in \mathbb{R}^{N\times W\times H\times D}$. Then these descriptors are fed into seven 3D convolution layers with channels 32, 32, 64, 64, 128, 128, 128. To predict $\Phi(\mathbf{x})$, three-layer MLP with channels 512, 256, 1 is employed; for $h(\mathbf{x})$, three-layer MLP with channels 512, 256, 128 is employed, and the embedding dimension is 128. These two branches share the first two layers.

For WGAN-GP network, we use five *conv1d* layers (with channels 512, 256, 128, 64, 1) and a *max-pooling* layer for the critic function D. The gradient penalty coefficient $\lambda = 1$.

For the decoder, we leverage a similar structure with TopNet [33]. Specifically, the decoder has 6 levels and each MLP in the decoder tree generates a small node feature embedding of size 8. When generating $N = 2048$ points, the root node has 4 children and all other internal nodes in subsequent level generate 8 children. Each MLP in the decoder is a has 3 stages with 256, 64, and 8 channels respectively.

Hyperparameters and Training For ShapeNet models, we choose $\eta_1 = 10.$, $\eta_2 = 1$, $\eta_3 = 0.1$ through the validation set; for SMPL human body dataset, we choose $\eta_1 = 10.$, $\eta_2 = 1$, $\eta_3 = 0$. In all our experiments without specification, the Beta prior distribution is fixed with $\alpha = 0.01$ and $\beta = 0.05$. The parameters of the network are optimized using Adam [17], with learning rate 1e-4.

4. Experiments

4.1. Comparison with Human Annotated Keypoints

In this section, we compare detected keypoints with those human annotated ones, in order to see if there is any semantic correspondence among keypoints.

Dataset Two datasets are utilized: ShapeNet-chair keypoint and KeypointNet [37] dataset. ShapeNet-chair keypoint set is proposed by SyncSpecCNN [36], which consists of thousands of keypoints annotated on ShapeNet chairs by experts. KeypointNet annotates millions of keypoints on models from 16 object categories of ShapeNet. We evaluate on airplanes, chairs and tables for KeypointNet and on chairs for ShapeNet-chair keypoint dataset. For both datasets, we randomly split 75%, 10% and 15% for train, val and test.

Metric We evaluate the performance by mean Intersection over Union (mIoU). Intersection is counted if the geodesic distance of a detected keypoint to its closest ground-truth is less than some geodesic threshold. Union is simply the union of detected and ground-truth keypoints.

Evaluation and Results We compare UKPGAN with USIP [20], D3Feat [2], Harris-3D [29], ISS [41] and SIFT-3D [27]. Training is done independently for each category. UKPGAN, USIP and D3Feat output keypoint probabilities which are refined through Non-Maximum-Suppression (NMS) with radius 0.1 and threshold ($p = 0.5$). Quantitative results are given in Figure 4. We see that UKPGAN aligns better with human annotated keypoints. It achieves much higher IoU than other methods. Qualitative visualizations are shown in Figure 3. UKPGAN gives keypoints that are intra-class consistent and edge/corner salient.

4.2. Detecting Stable Keypoints with Semantics under Different Human Poses

Dataset Skinned Multi-Person Linear model (SMPL) is a skinned vertex-based model that accurately represents a wide variety of body shapes in natural human poses. Human poses are controlled with three parameters, and we generate training data on the fly by changing these parameters. 2048 points are sampled uniformly from the original mesh.

Metric SMPL provides point-to-point correspondence across different human models. Given a pair of models (Model A, B), we evaluate detectors' stability and consistency by calculating Intersection of Union (IoU). An intersection is counted if a detected keypoint in Model A has its corresponding point in Model B detected too. The union is the summation of all the detected keypoints in both models. In order to take noise into account, consistency loss is also evaluated. It is calculated as the average distance between a detected point in Model A and its nearest detected neighbor in Model B, under ground-truth correspondences.

Evaluation and Results We evaluate the performance of UKPGAN method and compare it with USIP [20], D3Feat [2], Harris-3D [29], ISS [41] and SIFT-3D [27]. Keypoints are selected with threshold $p = 0.5$ with no NMS applied. Additionally, we adapt the number of predicted keypoints of baselines so that they are directly comparable to our model. More comparisons on fixed number of keypoints (10, 20, 40) with NMS enabled are given in our supplementary. Quantitative results are given in Table 1. Our method achieves the best IoU and consistency loss, suggesting it is robust and stable. Qualitative results are shown in Figure 8.

	IoU (%) ↑	Consis. ($\times 10^{-3}$) ↓
USIP [20]	23.9	4.6
D3Feat [2]	20.3	3.8
Harris-3D [29]	8.1	3.2
ISS [41]	8.1	3.3
SIFT-3D [27]	8.2	3.3
Ours	**66.6**	**1.2**

Table 1. **IoU (%) and Consistency Loss ($\times 10^{-3}$) results for SMPL dataset.** Our keypoint detector is stable under different deformations.

4.3. Keypoints for Real-World Registration

Dataset 3DMatch dataset [39] is an indoor registration benchmark. The test set contains 8 scenes with partially overlapped point cloud fragments and their corresponding transformation matrices. ETH dataset [25] is another outdoor registration benchmark, whose test set contains 4 scenes with overlapped fragments. Our keypoint detector is trained on ShapeNet dataset and then directly applied to 3DMatch and ETH. We down-sample 3DMatch and ETH point clouds using a voxel grid filter of size 0.03m and 0.02m, respectively.

Metric Geometric registration usually consists of two stages: keypoint detection and descriptor extraction. In order to compare the performance of different keypoint detectors, we leverage two state-of-the-art descriptors: PerfectMatch [10] and D3Feat [2], and combine them with our detector. As a baseline, random sampling, traditional detectors (i.e., ISS, SIFT-3D) or task-specific learning based detectors (i.e., D3Feat) are also evaluated. For each fragment, different number of points are given (i.e., 2500, 1000, 500, 250, 100) as budgets. We use NMS on keypoint score for D3Feat and UKPGAN, and random sampling is leveraged for traditional detectors to fulfill the budget requirement. We follow

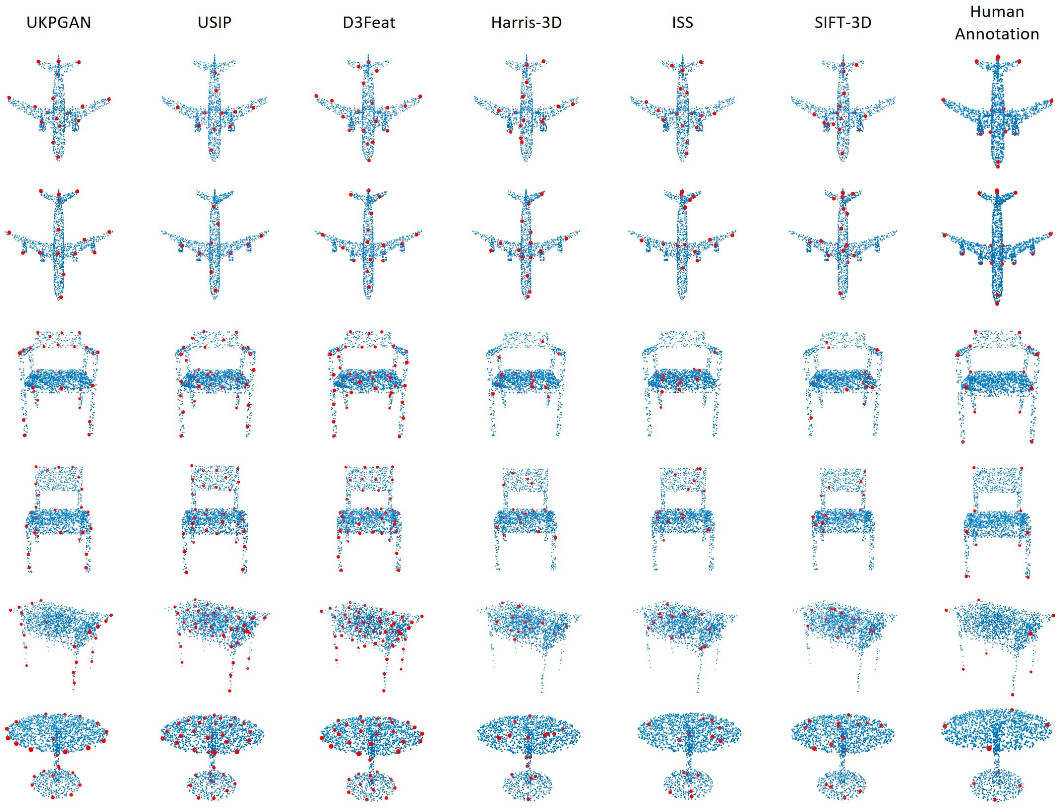

Figure 3. **Visualizations of six algorithms on unsupervised keypoint detection on ShapeNet models.**

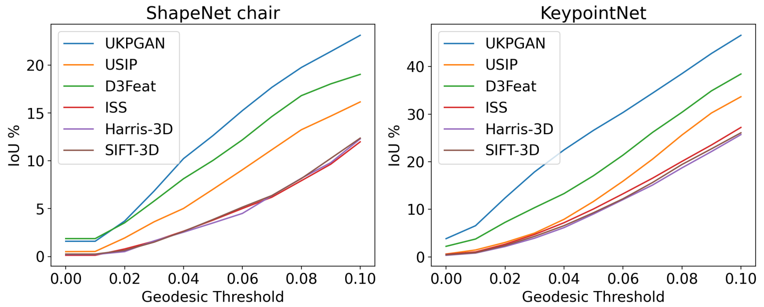

Figure 4. **mIoU results on ShapeNet chair dataset and KeypointNet.**

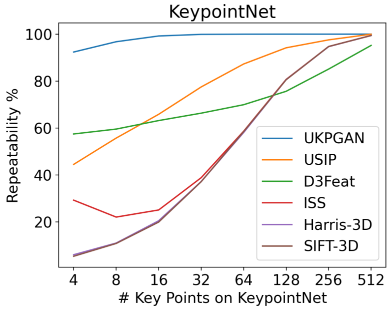

Figure 5. **Rotation repeatability.**

D3Feat to use Feature Matching Recall, Registration Recall and Inlier Ratio for evaluation. Feature Matching Recall is the percentage of successful alignment whose inlier ratio is above some threshold (i.e., $\tau_2 = 5\%$), which measures the matching quality of pairwise registration. A point-pair alignment is deemed successful if their distance is within some threshold (i.e., $\tau_1 = 0.1m$). Registration Recall is the percentage of successful alignment whose transformation error is below some threshold (i.e., RMSE $< 0.2m$), which better reflects the final performance. We use RANSAC with 50,000 max iterations to estimate the transformation matrices.

Evaluation and Results Results are shown in Table 2 and 3. The PerfectMatch and D3Feat descriptors are based on the pretrained model released by the authors. For PerfectMatch descriptor, our keypoint detector outperforms other detectors by a large margin, especially when the number of keypoints is small. For D3Feat descriptor, though D3Feat detector performs the best, the detector is trained together with the descriptor on real-world training data, while our keypoint detector is trained on synthetic ShapeNet models only. Besides, our method also outperforms other traditional keypoint detectors by a large margin. Our model could generalize to real-world data and may improve registration results.

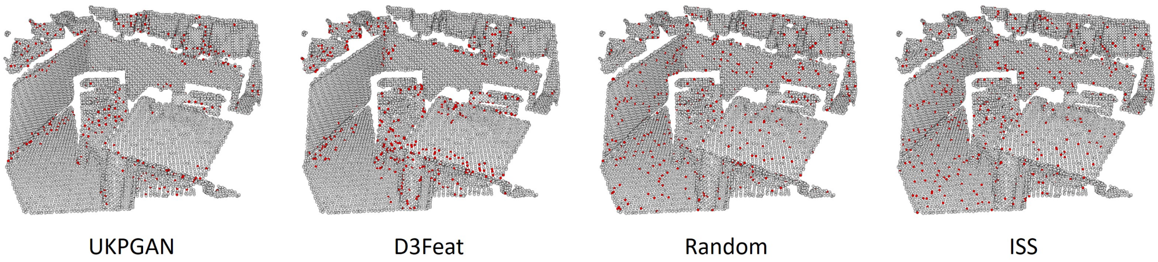

Figure 6. **Keypoint detection comparisons on 3DMatch Dataset.** Notice that our method is trained on synthetic models only, and achieve competitive results with that trained on real scenes (i.e., D3Feat). UKPGAN is able to give distinguishable keypoints for registration.

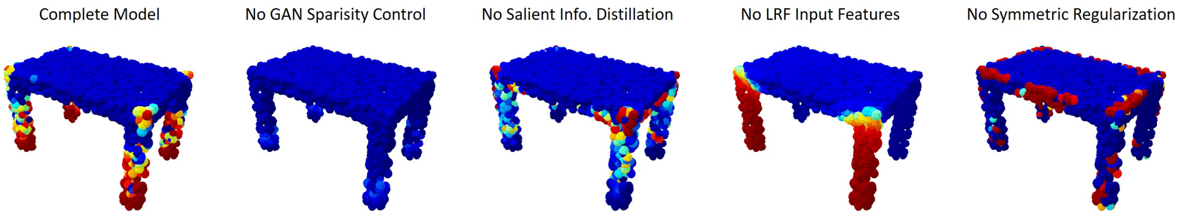

Figure 7. **Visualizations of ablation study on a ShapeNet table.** Colors indicate keypoint probabilities (red means high and blue means low). We see that without GAN sparsity control, our model fails to give meaningful keypoints.

Detector	Descriptor	Feature Matching Recall (%)					Registration Recall (%)					Inlier Ratio (%)				
		2500	1000	500	250	100	2500	1000	500	250	100	2500	1000	500	250	100
ISS [41]	PerfectMatch	90.3	**87.9**	82.7	71.9	51.7	75.5	68.6	55.4	37.9	14.0	28.1	22.6	18.7	15.7	12.9
SIFT [27]	PerfectMatch	90.3	87.7	82.5	74.5	52.7	**77.4**	68.1	56.4	35.8	11.7	28.0	22.6	18.5	15.1	12.4
Random	PerfectMatch	**90.4**	86.8	82.3	71.2	53.5	76.8	68.9	54.8	36.2	16.1	28.2	22.8	18.5	15.1	12.3
Ours	PerfectMatch	90.1	87.8	**85.6**	**83.1**	**74.2**	76.1	**72.5**	**70.0**	**63.6**	**37.6**	**28.5**	**25.4**	**25.7**	**24.5**	**18.8**
ISS [41]	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 [27]	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
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
D3Feat [2]	D3Feat	**95.5**	**94.5**	**94.1**	**93.1**	**90.6**	**84.3**	**83.6**	**82.5**	**78.1**	**67.2**	**40.5**	**42.6**	**44.0**	**44.7**	**45.6**
Ours	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

Table 2. **Registration result on 3DMatch.** We evaluate on two state-of-the-art descriptors, combined with different keypoint detectors.

Detector	Descriptor	Feature Matching Recall (%)					Registration Recall (%)					Inlier Ratio (%)				
		2500	1000	500	250	100	2500	1000	500	250	100	2500	1000	500	250	100
ISS [41]	PerfectMatch	59.1	41.3	23.1	11.3	6.3	48.8	28.1	12.8	5.2	1.3	11.4	9.0	8.0	6.8	6.8
SIFT [27]	PerfectMatch	58.5	39.4	23.9	10.8	6.2	45.5	26.9	12.9	6.0	0.9	11.3	9.0	7.5	6.8	6.5
Random	PerfectMatch	60.8	39.7	22.2	13.7	4.4	50.1	30.7	16.6	4.3	0.4	11.3	9.1	7.6	6.8	6.4
Ours	PerfectMatch	**68.1**	**62.4**	**53.6**	**44.8**	**29.6**	**58.2**	**45.5**	**32.3**	**19.1**	**6.1**	**18.7**	**16.2**	**14.2**	**11.8**	**10.0**
ISS [41]	D3Feat	37.9	24.4	16.3	10.8	6.2	25.6	18.1	8.9	4.7	1.7	8.8	7.7	7.2	6.6	7.5
SIFT [27]	D3Feat	36.8	24.6	14.9	10.2	5.5	28.4	16.7	9.0	3.0	1.1	8.7	7.7	7.0	7.2	6.7
Random	D3Feat	27.7	16.7	7.7	3.6	2.1	20.4	11.0	7.0	1.5	1.5	8.1	6.7	6.5	6.3	6.3
D3Feat [2]	D3Feat	**48.5**	**54.5**	**57.0**	**57.3**	**49.9**	**29.2**	**28.7**	**29.5**	**22.8**	**11.2**	10.9	**12.0**	**13.0**	**13.5**	**13.9**
Ours	D3Feat	47.5	43.1	37.4	33.0	21.5	28.3	22.0	14.2	10.9	3.9	**12.4**	11.6	10.9	9.9	9.2

Table 3. **Registration result on ETH.** We evaluate on two state-of-the-art descriptors, combined with different keypoint detectors.

Qualitative results are given in Figure 6.

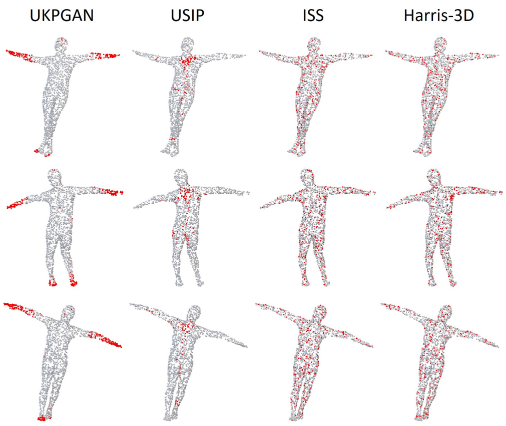

Figure 8. **Keypoint prediction results on SMPL dataset.**

4.4. Repeatability under Arbitrary Rotations

A good keypoint detector should be invariant to rotations, since orientations are often unknown. Therefore, rotation repeatability is an important metric to measure the quality of a keypoint detector. We evaluate on the test split of KeypointNet dataset, averaged over airplane, chair and table.

We follow the relative repeatability metric proposed in USIP as the evaluation metric. Given two point clouds of the same object, a keypoint in the first point cloud is considered repeatable if its distance to the nearest keypoint in the second point cloud is less than 0.1, under ground-truth transformations. We report the percentage of repeatable keypoints when different number of keypoints are detected.

We compare UKPGAN with USIP [20], D3Feat [2], Harris-3D [29], ISS [41] and SIFT-3D [27]. We generate 4, 8, 16, 32, 64, 128, 256, 512 most salient keypoints and calculate the relative repeatability respectively. The relative repeatability under arbitrary rotations on KeypointNet dataset is shown in Figure 5. Thanks to the local reference frame (LRF) extracted in our method, we achieve much higher keypoint repeatability than all previous methods. Even only four keypoints are detected, we achieve nearly 100% repeatability.

4.5. Ablation Study

In this section, we validate our design choices by conducting several ablation studies. Evaluation results are done on KeypointNet test split. Both IoU and rotation repeatability are evaluated. IoU is reported by NMS under threshold 0.1 and rotation repeatability is reported with 4 most salient keypoints. Quantitative and qualitative results are shown in Table 4 and Figure 7.

	IoU (%)			Rotation Rep. (%)		
	Airplane	Chair	Table	Airplane	Chair	Table
Ours	**68.8**	**36.2**	**34.7**	98.3	88.3	90.6
Ours w/o GAN Sparsity	36.3	27.2	23.1	**99.8**	**95.9**	99.6
Ours w/o Salient Info. Distill.	51.8	33.3	19.0	94.1	87.5	**99.9**
Ours w/o LRF Feat.	22.4	16.0	21.2	15.4	4.9	0.7
Ours w/o Symmetric Reg.	54.9	20.0	22.3	85.1	77.0	73.0

Table 4. **Results of various ablation studies.**

GAN-based Keypoint Sparsity Control. GAN allows learning keypoint distributions with easily controllable parameters. We experienced with L1 norm and found that it fails to output a meaningful keypoint distribution by tuning the coefficient of norm loss, as shown in Figure 7.

Salient Information Distillation. Salient information Distillation is another important module for our model. We compare our complete model with a baseline that implements a simple averaging instead of max-pooling. It shows that with no salient information distillation, salient parts of models are not detected.

Local Rotation Invariant Descriptors. Local rotation invariant descriptors play an important role in maintaining repeatability under arbitrary rotations. If we replace it with raw XYZ features, both IoU and rotation repeatability drop.

Symmetric Regularization. In Section 3.5, we integrate a symmetric invariance prior into our model, which is helpful since the extracted descriptors are only rotation invariant rather than symmetric invariant. If we remove symmetric regularization, we see that detected keypoints are not symmetric anymore in Figure 7.

5. Conclusion

In this work, we proposed a keypoint detector which could detect meaningful points in an unsupervised way. The key contributions of our method are GAN-based sparsity control and salient information distillation modules. Experiments show that our UKPGAN detector can produce stable points on rigid and non-rigid objects. Moreover, our method also generalizes well to real scenarios.

6. Acknowledgements

This work was supported by the National Key Research and Development Project of China (No. 2021ZD0110700), the National Natural Science Foundation of China under Grant 51975350, Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Shanghai Qi Zhi Institute, and SHEITC (2018-RGZN-02046). This work was also supported by the Shanghai AI development project (2020-RGZN-02006) and "cross research fund for translational medicine" of Shanghai Jiao Tong University (zh2018qnb17, zh2018qna37, YG2022ZD018).

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017. 4

[2] 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. 1, 2, 5, 7, 8

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006. 2

[4] M Bueno, J Martínez-Sánchez, H González-Jorge, and H Lorenzo. Detection of geometric keypoints and its application to point cloud coarse registration. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 41, 2016. 1

[5] Umberto Castellani, Marco Cristani, Simone Fantoni, and Vittorio Murino. Sparse points matching by combining 3d mesh saliency with statistical descriptors. In *Computer Graphics Forum*, volume 27, pages 643–652. Wiley Online Library, 2008. 1, 2

[6] Nenglun Chen, Lingjie Liu, Zhiming Cui, Runnan Chen, Duygu Ceylan, Changhe Tu, and Wenping Wang. Unsupervised learning of intrinsic structural representation points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9121–9130, 2020. 2

[7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 1

[8] 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. 2

[9] Georgios Georgakis, Srikrishna Karanam, Ziyan Wu, Jan Ernst, and Jana Košecká. End-to-end learning of keypoint detector and descriptor for pose invariant 3d matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1965–1973, 2018. 2

[10] Zan Gojcic, Caifa Zhou, Jan D Wegner, and Andreas Wieser. The perfect match: 3d point cloud matching with smoothed densities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5545–5554, 2019. 3, 5

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 3

[12] Christopher G Harris, Mike Stephens, et al. A combined corner and edge detector. In *Alvey vision conference*, volume 15, pages 10–5244. Citeseer, 1988. 1

[13] Tong He, Haibin Huang, Li Yi, Yuqian Zhou, Chihao Wu, Jue Wang, and Stefano Soatto. Geonet: Deep geodesic networks for point cloud analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6888–6897, 2019. 1

[14] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *Advances in neural information processing systems*, pages 4016–4027, 2018. 2

[15] Tomas Jakab, Richard Tucker, Ameesh Makadia, Jiajun Wu, Noah Snavely, and Angjoo Kanazawa. Keypointdeformer: Unsupervised 3d keypoint discovery for shape control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12783–12792, 2021. 2

[16] Marc Khoury, Qian-Yi Zhou, and Vladlen Koltun. Learning compact geometric features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 153–161, 2017. 2

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 4

[18] Chang Ha Lee, Amitabh Varshney, and David W Jacobs. Mesh saliency. *ACM transactions on graphics (TOG)*, 24(3):659–666, 2005. 1, 2

[19] Jiaxin Li, Ben M Chen, and Gim Hee Lee. So-net: Self-organizing network for point cloud analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9397–9406, 2018. 1

[20] Jiaxin Li and Gim Hee Lee. Usip: Unsupervised stable interest point detection from 3d point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 361–370, 2019. 1, 2, 5, 8

[21] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 1, 2

[22] Ajmal S Mian, Mohammed Bennamoun, and Robyn Owens. Three-dimensional model-based object recognition and segmentation in cluttered scenes. *IEEE transactions on pattern analysis and machine intelligence*, 28(10):1584–1601, 2006. 1

[23] Zak Murez, Tarrence van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *European Conference on Computer Vision*, pages 414–431. Springer, 2020. 1

[24] John Novatnack and Ko Nishino. Scale-dependent 3d geometric features. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 1

[25] François Pomerleau, Ming Liu, Francis Colas, and Roland Siegwart. Challenging data sets for point cloud registration algorithms. *The International Journal of Robotics Research*, 31(14):1705–1711, 2012. 5

[26] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition, pages 652–660, 2017. 4

[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. 1, 2, 5, 7, 8

[28] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 1, 2

[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, 2011. 1, 2, 5, 8

[30] 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. 1, 2

[31] Minhyuk Sung, Hao Su, Ronald Yu, and Leonidas J Guibas. Deep functional dictionaries: Learning consistent semantic structures on 3d models from functions. In *Advances in Neural Information Processing Systems*, pages 485–495, 2018. 2

[32] Supasorn Suwajanakorn, Noah Snavely, Jonathan J Tompson, and Mohammad Norouzi. Discovery of latent 3d keypoints via end-to-end geometric reasoning. In *Advances in neural information processing systems*, pages 2059–2070, 2018. 2

[33] Lyne P Tchapmi, Vineet Kosaraju, Hamid Rezatofighi, Ian Reid, and Silvio Savarese. Topnet: Structural point cloud decoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 383–392, 2019. 4

[34] Hanyu Wang, Jianwei Guo, Dong-Ming Yan, Weize Quan, and Xiaopeng Zhang. Learning 3d keypoint descriptors for non-rigid shape matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 1

[35] J Gerard Wolff. Information compression as a unifying principle in human learning, perception, and cognition. *Complexity*, 2019, 2019. 2

[36] Li Yi, Hao Su, Xingwen Guo, and Leonidas J Guibas. Syncspeccnn: Synchronized spectral cnn for 3d shape segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2282–2290, 2017. 2, 5

[37] 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. 5

[38] Maciej Zamorski, Maciej Zięba, Piotr Klukowski, Rafał Nowak, Karol Kurach, Wojciech Stokowiec, and Tomasz Trzciński. Adversarial autoencoders for compact representations of 3d point clouds. *Computer Vision and Image Understanding*, 193:102921, 2020. 3

[39] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1802–1811, 2017. 5

[40] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2694–2703, 2018. 2

[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. 1, 2, 5, 7, 8