Learning Local Descriptors with Adversarial Enhancer from Volumetric Geometry Patches

Jing Zhu
NYU Multimedia and Visual Computing Lab
Dept. of CSE, NYU Tandon School of Engineering, USA
Dept. of ECE, NYU Abu Dhabi, UAE

Yi Fang*
NYU Multimedia and Visual Computing Lab
Dept. of CSE, NYU Tandon School of Engineering, USA
Dept. of ECE, NYU Abu Dhabi, UAE

ABSTRACT

Local matching problems (e.g. key point matching, geometry registration) are significant but challenging tasks in computer vision field. In this paper, we propose to learn a robust local 3D descriptor from volumetric point patches to tackle the local matching tasks. Intuitively, given two inputs, it would be easy for a network to map the inputs to a space with similar characteristics (e.g. similar outputs for similar inputs, far different outputs for far different inputs), but the difficult case for a network would be to map the inputs into a space with opposite characteristics (e.g. far different outputs for very similar inputs but very similar outputs for far different inputs). Inspired by this intuition, in our proposed method, we design a siamese-network-based local descriptor generator to learn a local descriptor with small distances between match pairs and large distances between non-match pairs. Specifically, an adversarial enhancer is introduced to map the outputs of the local descriptor generator into an opposite space that match pairs have the maximum differences and non-match pairs have the minimum differences. The local descriptor generator and the adversarial enhancer are trained in an adversarial manner. By competing with the adversarial enhancer, the local descriptor generator learns to generate a much stronger descriptor for given volumetric point patches. The experiments conducted on real-world scan datasets, including 7-scenes and SUN3D, and the synthetic scan augmented ICL-NUIM dataset show that our method can achieve superior performance over other state-of-the-art approaches on both keypoint matching and geometry registration, such as fragment alignment and scene reconstruction.

CCS CONCEPTS

• **Theory of computation → Adversarial learning**; • **Computing methodologies → Matching**; **Neural networks**;

KEYWORDS

keypoint matching; geometry registration; adversarial learning

*Corresponding Author (Email: yfang@nyu.edu)

MM '18, October 22–26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240666

ACM Reference Format:
Jing Zhu and Yi Fang. 2018. Learning Local Descriptors with Adversarial Enhancer from Volumetric Geometry Patches. In *2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3240666

1 INTRODUCTION

A good local geometric descriptor enables a wide range of applications, such as semantic segmentation, point matching, and scene reconstruction. Though many methods have been proposed to learn descriptors from multiple formats of 3D data, such as 3D mesh, 3D scan data – RGB-D images, volumetric point patches, most of them focused on learning a global descriptor. However, how to obtain a good local descriptor remains a challenging but interesting computer vision task. In this paper, we aim to learn a robust local 3D descriptor that can be used as representations to match the sample points in the fragments, and then we can compute the rigid transformation matrix (including rotations, scales, translations) between matching points in any two fragments, finally align the two fragments using the computed transformation (as shown in Figure 1).

At the beginning of the descriptors research history, the early researchers were dedicated in designing hand-crafted descriptors (features) that could well describe the 3D objects, such as spin-image [22] and FPFH [33]. In the most recent decade, with the development of the deep learning techniques, there have been a variety of applications using powerful convolutional neural networks (mostly) in 2D computer vision area, e.g. image object classification, image scene parsing, image translation. Meanwhile, researchers have shown their great interests in learning a robust local descriptor for 3D objects using convolutional neural networks. However, different from the 2D color images that consist of only pixel values in three channels, the format of the 3D object is much more complicated, which is usually represented as vertexes coordinates and triangles connected by vertexes. It is nearly impossible to apply CNN techniques directly on such format of 3D objects to solve 3D computer vision problems. Inspired by the successful 2D CNN applications, some recent approaches proposed to first render (multi-view) images from 3D objects, then learned a local 3D descriptor based on the rendered images. In this way, those researchers were able to utilize the 2D CNN framework directly to learn a 3D descriptor. However, their performance significantly depended on the quality of the rendered images. What is more, for the approaches using multi-view rendered images, they usually required more than 10 rendered images for each 3D model or each local part on a model, which makes it very difficult to be generalized or applied to other 3D objects, datasets or applications.

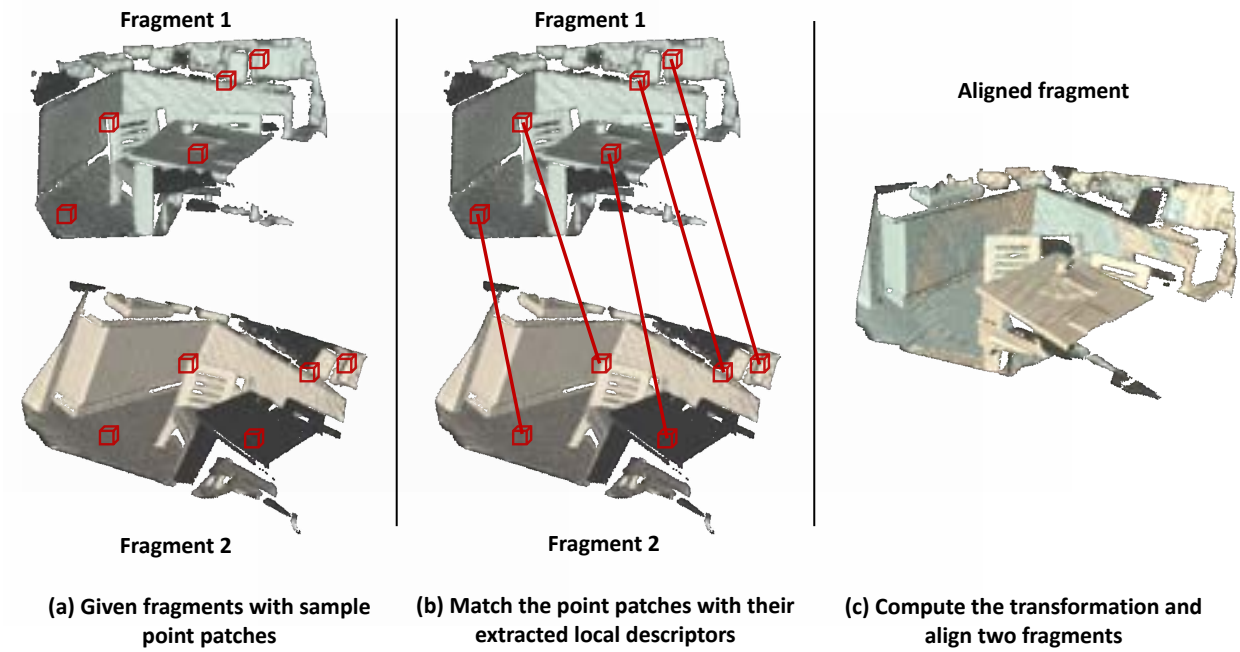

Figure 1: The pipeline of our work. Given pairs of points with their local volumetric patches, we aim to learn a model that is able to generate robust local descriptors for key point matching between all the given point pairs. Moreover, we can compute the rigid transformation between the match pairs of points of two fragments. Finally, the two fragments can be further aligned using the computed transformation.

Recently, the popularization of 3D volumetric models provides us another feasible direction to learn a robust 3D local descriptor. One of the advantages of working on the volumetric models is that it enables us to design a 3D CNN architecture directly on 3D volumetric models. 3D-GAN [43] and SSCNet [40] are both latest works based on volumetric models in 3D computer vision area for model generation and scene completion, respectively. 3DMatch [46] was the first published work that learned a local descriptor on 3D volumetric patches generated from RGB-D frames. Though their work was promising, their model was trained on more than 10 million correspondences for more than 8 days, which was super time-consuming, hard to generalize and not efficient. The causes might be 1) the unnecessary deep network structure and 2) too complex training strategy. In this paper, we are seeking not only a more effective but also a more efficient approach to learn a robust local geometric descriptor on the volumetric point patches that contain the truncated distance function (TDF) values computed from the points (and their neighbors) on the RGB-D scene frames.

On the other hand, siamese networks with contrastive loss have been proved their effectiveness on matching problems, such as shape matching, RGB-D image matching, cross-domain matching and point matching. Admiring the superior learning power of the siamese network with contrastive loss, we create our local descriptor generator with a pair of siamese deep 3D convolutional neural networks. Moreover, inspired by the recent advances of the generative adversarial networks (GANs), we consider to train a local descriptor generator in adversarial manner so that the training process could be improved. Motivated by the idea that a robust learned descriptor is difficult to be mapped into an opposite feature space, we introduce an siamese-network-based adversarial enhancer with an opposite loss to the local descriptor generator. That is to say, during the training process, the adversarial enhancer is learning to minimize the distances between the generator-learned features for non-match pairs while to maximize the distances for match pairs. To compete with the enhancer and win the two-player game, it enforces the local descriptor generator to learn a robust and powerful descriptor for giving volumetric point patches with very small distances between the learned features for match pairs but very large differences between learned features for the non-match pairs. As a result, the performance of the local descriptor generator can be improved.

To validate the performance of our proposed method, we conduct experiments on the dataset that is constructed from SUN3D dataset [19, 44], 7-Scenes dataset [39], BundleFusion dataset [11] and the augmented ICL-NUIM dataset for two different tasks, basic keypoint matching and geometric registration with fragment alignment and scene reconstruction on the computed volumetric point patches. We report the quantitative analysis of the keypoint matching problem. For the geometric registration task, we provide quantitative results for fragment alignment on real-world scan scenes and synthetic scenes separately. We also visualize some fragment alignment results and scene reconstructions using aligned fragments for qualitative evaluation. The experimental results demonstrate

that our proposed method is able to learn a robust local geometric descriptor on the volumetric point patches with superior performance over the state-of-the-art methods with a large gap on both the keypoint matching and geometry registration tasks. In order to verify the effectiveness of the adversarial enhancer, we present the performance using our designed local descriptor generator without the adversarial enhancer in all the experiments for comparison. The performance gap implies that the adversarial enhancer can significantly improve the learning ability of the local descriptor generator.

In summary, the main contributions of our work can be concluded as:

- To address the challenging local geometry matching problem, we propose to learn an improved siamese-network-based local descriptor generator on volumetric point patches with a siamese-network-based enhancer network.
- Specially, our adversarial enhancer is optimized with the opposite loss function to the local descriptor generator. In order to compete the adversarial enhancer, the local descriptor generator is enforced to robustly learn more similar features for match pairs while far different features for non-match pairs.
- The experiment results demonstrate that our proposed framework can generate effective 3D local descriptor that can achieve superior performance over the state-of-the-art methods for keypoint matching, geometry registration (in fragment alignment and scene reconstruction) on the point clouds fragments from real-world scan SUN3D dataset, 7-Scenes dataset and synthetic ICL-NUIM dataset.

Our paper is organized as follows: in Section 2, we review some recent works and concepts related to our work, such as hand-crafted 3D local descriptors, learned 3D local descriptors and some GANs applications on 3D computer vision field. In Section 3, we describe the pipeline of our proposed method and details of our network structure. In Section 4, we quantitively evaluate our learned local descriptor on keypoint matching and geometry registration tasks, and provide the qualitative analysis with visualization of some aligned fragments and scenes reconstructed from multiple fragments. Finally, we conclude our work in Section 5.

2 RELATED WORK

In this section, we briefly review some previous works related to ours, including some classic and popular approaches that extracted hand-crafted local feature, recent methods that learned local 3D representations using deep learning techniques, introduction of adversarial learning using neural networks (e.g. GANs) and some successful applications with the generative adversarial networks (GANs) technique.

2.1 Hand-crafted 3D Local Descriptors

In the early attempts to find a representation for 3D local parts, researchers started their exploration from designing hand-crafted descriptors upon the surface geometry structures of 3D mesh. One of the well-known traditional methods is the spin image one [22], where Johnson et al. proposed to create spin images at oriented points of 3D meshes, and then used a set of spin images as representation for 3D object recognition. Another approach [2] captured the coarse distributions from the sampled points to the remaining points as the local descriptor to find the point correspondences of the 3D object. Observing that geodesic distance worked well on the problems dealing with graph-like structures, Zhang et al. [47] got the distribution of the average geodesic distance between points over the meshes as the descriptors. In addition, curvature values [14] and shape diameter [38] were two important considerations when designing a local 3D descriptor. Most of the recent proposed hand-crafted 3D descriptors are histogram-based descriptors [1, 6, 32, 41]. For example, Rusu et al. bined the neighborhood's geometrical properties into a histogram as point feature [34, 35], and later improved their descriptor by fastening the computation process with some optimizations called Fast point feature histograms [33].

Though the above hand-crafted descriptors provided inspiration for 3D local descriptor research, they all have the limitation that their methods only worked on synthetic 3D meshes represented with vertexes coordinates and triangles between them. Nowadays, many 3D objects are captured from real-world using low-cost sensors, and they are usually represented as different data format, such as point clouds, RGB-D images. However, it is not easy to apply the hand-crafted descriptors on the captured 3D objects due to the lack of obvious surface geometry structures in the objects. In our work, we focus on designing a local 3D descriptor from the volumetric point patches. Since volumetric point patches can be easily computed from RGB-D images, depth images, volumetric models or point clouds, our descriptor could be conveniently extended to the local matching problems on those kinds of real-world 3D data.

2.2 Learned 3D Local Descriptors

On the other hand, inspired by the great success of applying deep learning techniques on various applications in graphic and vision community, such as retrieval, classification and transfer learning [7, 10, 12, 17, 20, 23, 25, 37, 49], many researchers have tried to develop a local 3D descriptor using deep learning techniques in their recent works. Since the mesh format of a synthetic 3D object (represented as vertexes coordinates and triangles) is difficult to be directly fed into a traditional convolutional neural network, most of current exisxting works first converted the synthetic 3D objects into another format, such as extracted hand-crafted features, rendered images before using the convolutional neural networks. For example, Guo et al. [18] fed a set of hand-crafted features extracted from triangle faces of meshes into 2D CNNs to learn a local geometric descriptor for each triangle face labelling. In paper[45], Yi et al. learned a local descriptor on the spectral domains of shapes for key point prediction and 3D shape part segmentation. In another work, Huang et al. [21] proposed to learn a local 3D descriptor on multi-scale images rendered from points on the 3D objects. In addition, some published works discussed their approaches on the manifolds of the deformable models, especially on human bodies [3, 4, 28, 30].

According to our knowledge, 3DMatch [46] is the first work proposed to learn a local descriptor on volumetric point patches that contained truncated distance function (TDF) values computed from the local part (points and their neighbors) of the real-world

3D data (e.g. depth scans). While their method showed impressive results in local matching problems, their training process was very time-consuming due to the unnecessary deep network structure and complex training strategy. In our work, we are seeking a way to alleviate the training process of learning a local descriptor but with equal or even better matching performance. Admiring the advantages of using volumetric point patches and the learning power of siamese network structure in matching problems, we build the framework on our designed siamese network structure that takes the volumetric point patches as inputs and output their learned local descriptors.

2.3 GAN-based Descriptors

Introducing by Goodfellow et al. [15], generative adversarial networks (GANs) have been proved its excellent learning ability on multiple tasks in computer vision area. A classic structure of generative adversarial networks (GANs) consist of one generator G and one discriminator D, and both are multilayer neural networks. The generator and discriminator are usually trained as a two-player minimax game with a competing loss. During the training process, the generator is learning to synthesize data as "real" as the real data, while the discriminator is learning to improve its distinguish ability of the real data. Optimization of the competing loss function can be acquired when the generative data distribution is equal to the real data distribution.

Most of the current research works applied GANs techniques to address tasks on 2D computer vision field, such as image generation, image domain adaption, image-to-image translation and image completion [5, 26, 27, 50]. We also realize the potential of using adversarial strategy on feature learning tasks. As one of the attempts, Radford et al. [31] extracted max-pooling features from learned discriminator as the unsupervised features for image classification. [43] and [48] are the recent works to apply GANs technique on the volumetric 3D model generation, where they obtained the unsupervised features from multiple layers of their trained discriminator as the representations for 3D model classification.

Besides extracting features directly from the discriminators of GANs framework, some researchers extended the adversarial learning idea by adding a discriminator into some classical feature learning models to improve the discriminability of the original descriptors. For example, Wang et al. [42] challenged an image classifier by introducing two adversarial networks to generate some rare samples that were difficult for normal classifiers to recognize, which in turn significantly improved the image classifier. One more case, Salimans et al. [36] successfully improved the classified ability of the image features by adding some samples generated from a adversarial generator. In paper [16], the authors also highlighted the importance of adding adversarial samples when training a classifier. In our work, we explore the effectiveness of adding an adversarial siamese-network-based enhancer when learning a local 3D descriptor.

3 APPROACH

In this section, we provide details of our method for learning a robust local descriptor on volumetric point patches. We start with brief description of the basic structure and concepts of general siamese network followed by the presentation for our proposed framework architecture.

3.1 Siamese Deep Convolutional Networks

Introduced by [9] with applications on face verification, siamese deep convolutional network has been proved its great ability to learn a descriptor in a wide range of areas. Generally, the siamese networks are a pair of deep convolution networks with exactly same network architecture and weight sharing between them. Contrastive loss is one of the classic loss function used to train siamese networks. Given a pair of inputs in the same data format, the siamese networks map the input data into a common feature space, where the feature difference between the feature vectors is small if they are a match pair, and large if they are a non-match pair. Let $\boldsymbol{x_1}$, $\boldsymbol{x_2}$ be a pair of inputs, and label $y = 1$ means the pair is matched while $y = 0$ means the given pair is not matched. Then the siamese networks can be updated by minimizing the following loss function

$$L = \frac{1}{2N}\sum_{j=1}^{N} y * d_j^2 + (1-y) * \max(margin - d_j, 0)^2, \qquad (1)$$

where d denotes the feature differences (usually the Euclidean distance) between the two feature vectors on the learned feature space, N is the total number of training sample pairs, and $margin$ constrains the differences between non-match pairs. The optimization of the loss function above can be solved by applying classic back-propagation algorithm. The parameters in both networks will be updated with the same gradients in each training epoch.

3.2 Network Architecture

By introducing the adversarial enhancer, we aim to improve the local descriptor generator so that it can generate more robust local features for given volumetric point patches with smaller distances between the learned features for match pairs but larger distances between learned features for non-match pairs. Figure 2 shows the framework of our proposed method, which consists of a local descriptor generator and an adversarial enhancer. We will present the structure of these two parts in detail below.

Local Descriptor Generator Our local descriptor generator G is a deep siamese neural network that maps given input volumetric point patches into a feature space, where the distances between features of match pairs are small while the distances between features of non-match pairs are large. Each in the pair of siamese networks includes four 3D convolution layers with channel size $\{64, 128, 256, 512\}$, kennel size $3 \times 3 \times 3$ and stride 1, followed by a global max pooling layer and two fully-connected layers with neuron size $\{512, 256\}$. ReLU layer is added to connect each two layers except the last fully-connected layer. All of the network parameters are shared between the networks. The final 256-dimensional output vectors are the learned local descriptors for the given volumetric point pairs. We utilize contrastive loss as the basic loss for our local descriptor generator:

$$\begin{aligned} L_G = &\frac{1}{2N}\sum y * ||G(\boldsymbol{x_1}) - G(\boldsymbol{x_2})||^2 + \\ &(1-y) * \max(margin - ||G(\boldsymbol{x_1}) - G(\boldsymbol{x_2})||, 0)^2, \end{aligned} \qquad (2)$$

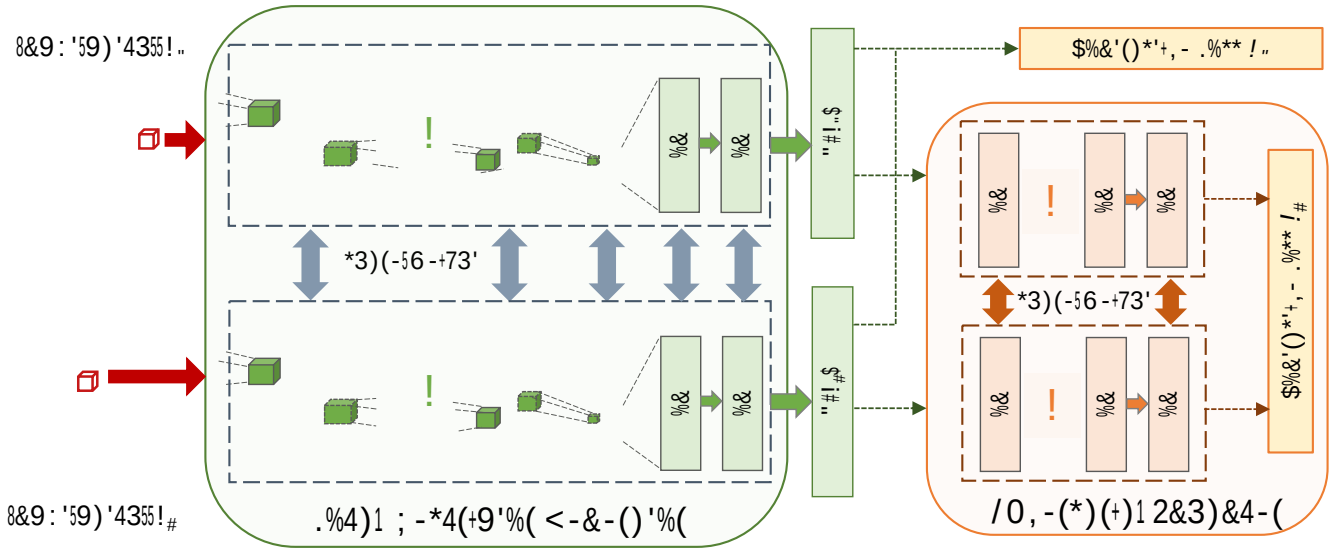

Figure 2: The framework of our proposed method. It consists of two parts: a local descriptor generator and an adversarial enhancer. The local descriptor generator contains a siamese deep convolutional neural networks, while the adversarial enhancer has a siamese networks with only simple fully-connected layers. Optimizing with the contrastive loss L_E which is opposite to the loss in the local descriptor, the enhancer is introduced to boost the generator in adversarial manner. In order to compete with the enhancer, the local descriptor generator must learn to generate much stronger feature with smaller difference between match pairs while larger difference between non-match pairs.

where $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ represent a pair of volumetric patches with size $30 \times 30 \times 30$, y denotes whether the given pair is a match or not (in our case, $y = 1$ for match pairs, $y = 0$ for non-match pairs). $||\cdot||$ is the Euclidean distance between the two vectors.

Adversarial Enhancer The purpose of the adversarial enhancer is to improve the learning ability of the local descriptor generator so that it could generate more similar outputs for match pairs but far more different outputs for non-match pairs. Ideally, the distances between descriptors of match pairs are 0 while the distances between the descriptors of non-match pairs are as large as possible. Motivated by the ideal case, we design an enhancer network that tries to maximize the distances between the learned descriptors of match pairs while minimize the distances between the learned descriptors of non-match pairs. If the learned descriptor is strong enough, it would be very difficult for the enhancer network to learn such kind of space.

The inputs for our adversarial enhancer are the outputs (local descriptors) from the local descriptor generator $G(\boldsymbol{x}_1)$ and $G(\boldsymbol{x}_2)$. We build our enhancer E with a pair of siamese networks so that it could be trained with the local descriptor generator simultaneously. The siamese network in the enhancer has four fully-connected layers with neuron size $\{256, 256, 128, 128\}$. ReLU layer is added between each two layers. Same as normal siamese networks, the network parameters are shared. Let $E(\cdot)$ be the 128-dimensional output of the enhancer network, then the loss function of the enhancer can be described as

$$\begin{aligned}\min_E L_E = &\frac{1}{2N}\sum (1-y) * ||E(G(\boldsymbol{x}_1)) - E(G(\boldsymbol{x}_2))||^2 + \\ &y * \max(margin - ||E(G(\boldsymbol{x}_1)) - E(G(\boldsymbol{x}_2))||, 0)^2.\end{aligned} \tag{3}$$

Network Training In order to learn a robust descriptor that cannot be easily mapped into the space with opposite characteristic, the local descriptor generator G learns to compete the adversarial enhancer with the adversarial loss

$$\begin{aligned}L_{G_{En}} = &\frac{1}{2N}\sum y * ||E(G(\boldsymbol{x}_1)) - E(G(\boldsymbol{x}_2))||^2 + \\ &(1-y) * \max(margin - ||E(G(\boldsymbol{x}_1)) - E(G(\boldsymbol{x}_2))||, 0)^2.\end{aligned} \tag{4}$$

Therefore, the loss function of the local descriptor generator can be extended as

$$\min_G L_{LDG} = L_G + \lambda * L_{G_{En}}. \tag{5}$$

We use ADAM optimizer to obtain the optimal network parameters with beta value $\beta = 0.5$, learning rate 0.0001 and *margin* 1.0. Considering that our enhancer here works like an auxiliary for the local feature generator, the loss $L_{G_{En}}$ should not overwhelm L_G. After several trials, we find that 0.01 is the most suitable value for λ to train the whole framework. The parameters in the local descriptor generator and the adversarial enhancer are updated separately in each epoch.

Table 1: Matching error comparisons on keypoint matching with state-of-the-art methods on the testing dataset constructed from the SUN3D and 7-scenes datasets.

Method	Matching Error (%)
Spin-Images [22]	83.7
FPFH [33]	61.3
3DMatch [46]	35.3
Ours without Adversarial Enhancer	**32.4**
Ours with Adversarial Enhancer	**29.5**

Table 2: The precision and recall comparisons of geometry registration with state-of-the-art methods on the synthetic scenes in the augmented ICL-NUIM dataset.

Method	Recall (%)	Precision (%)
Super 4PCS [29]	17.8	10.4
FPFH [33]	44.9	14.0
Variant FPFH [8]	59.2	19.6
FPFH [33] + RANSAC	46.1	19.1
Spin-Images [22] + RANSAC	52.0	21.7
3DMatch [46] + RANSAC, fine-tuned	65.1	25.2
Ours without Adversarial Enhancer + RANSAC	**58.6**	**25.3**
Ours with Adversarial Enhancer + RANSAC	**60.3**	**28.3**

Table 3: The precision and recall comparisons of geometry registration with state-of-the-art methods on real-world scan scenes constructed from the SUN3D and 7-scenes datasets.

Method	Recall (%)	Precision (%)
FPFH [33] + RANSAC	44.2	30.7
Spin-Images [22] + RANSAC	51.8	31.6
3DMatch [46] + RANSAC	66.8	40.1
Ours without Adversarial Enhancer + RANSAC	**69.1**	**40.5**
Ours with Adversarial Enhancer + RANSAC	**72.0**	**42.9**

4 EXPERIMENTS

To comprehensively validate our proposed framework, we conduct two different experiments on large-scale RGB-D reconstruction datasets, including keypoint matching and geometry registration. We present the experiment settings, quantitive matching analysis and qualitative registration results obtained by applying our proposed local descriptor on the computed volumetric patches from the indoor scenes. The experimental results demonstrate that our method can learn a robust representation for local 3D volumetric point patches to solve classic local matching problems. Moreover, though our model is trained on fewer samples for less time, it outperforms the state-of-the-arts methods with lower keypoint matching error and higher geometry registration precisions.

4.1 Keypoint Matching

In this task, we train our proposed framework on the point patches sampled from large-scale SUN3D dataset [19, 44], 7-Scenes dataset [39] and RGB-D Scenes V2 dataset [24]. There are totally more than $100K$ RGB-D frames categorized into 54 different indoor scenes, including offices, apartments, hotels and study-rooms. Following the setting in 3DMatch paper [46], we split the 54 scenes into two non-overlap training set and testing set, which contain 46 and 8 non-overlap scenes, respectively. For fair comparison, we use 1) the same sample strategy as the one used in 3DMatch to sample $30K$ pairs of $30 \times 30 \times 30$ volumetric point patches from the 46 training scenes to construct our own training set, where $15K$ are match pairs and the rest are non-match pairs; and 2) the same testing set as the one in 3DMatch, which contains $10K$ pairs of point patches sampled from the 8 test scenes with a ratio of $1:1$ for match pairs and non-match pairs.

After training our model on the constructed training set, for each point patch in the testing set, we extract 256-dimensional outputs from the trained local descriptor generator as their representations, and then match them based on the Euclidean distances calculated between the extracted learned features for each keypoint pair. We measure the matching performance using the false-positive rate (matching error), the lower the better. The matching error of our method is 29.5% when recall reaches 95% (as shown in Table 1). More importantly, our designed model only needs approximately 14 hours for training, which is nearly 14 times faster than 3DMatch (trained for more than 8 days).

Furthermore, we collect the publicly available results of state-of-the-art approaches from the 3DMatch website [1] for comparison, see Table 1. Though our method is only trained on a smaller dataset with $30K$ pairs of volumetric point patches, our method achieves the lowest matching error among all the compared approaches, e.g. spin-images [22], Fast Point Feature Histograms (FPFH) [33], 3DMatch [46]. We also report the keypoint matching performance (32.4% error) when using our framework without the adversarial enhancer with the same experimental settings, e.g. batch size, learning rate, epoch, etc. The improvement of the performance using the framework with the adversarial enhancer clearly demonstrates the effectiveness of the adversarial enhancer.

4.2 Geometry Registration

In addition to the keypoint matching task, we further evaluate our learned local descriptor on another local matching problem – geometry registration. Geometry registration is a challenging task that match two or more fragments which belong to the same scene as a whole one. Following the instruction of 3DMatch, we first extract the 256-dimensional local descriptors for all the sampled point patches (e.g. $5K$ sampled point pairs) on each fragment, and then apply the RANSAC [13] algorithm to find out the optimum transformation metric based on the match point pairs between two fragments. After that, we can align any two fragments (from the same scene) with the computed transformation, and reconstruct the scenes by combining multiple aligned fragments. If a method can generate more robust local features that well describe the local points, match point pairs would be detected easily, and as a consequence, the alignment results will be better.

To quantitatively verify the registration performance, we calculate the precisions and recalls among the transformations between each testing fragment pair, following the evaluation measurement

[1] http://3dmatch.cs.princeton.edu

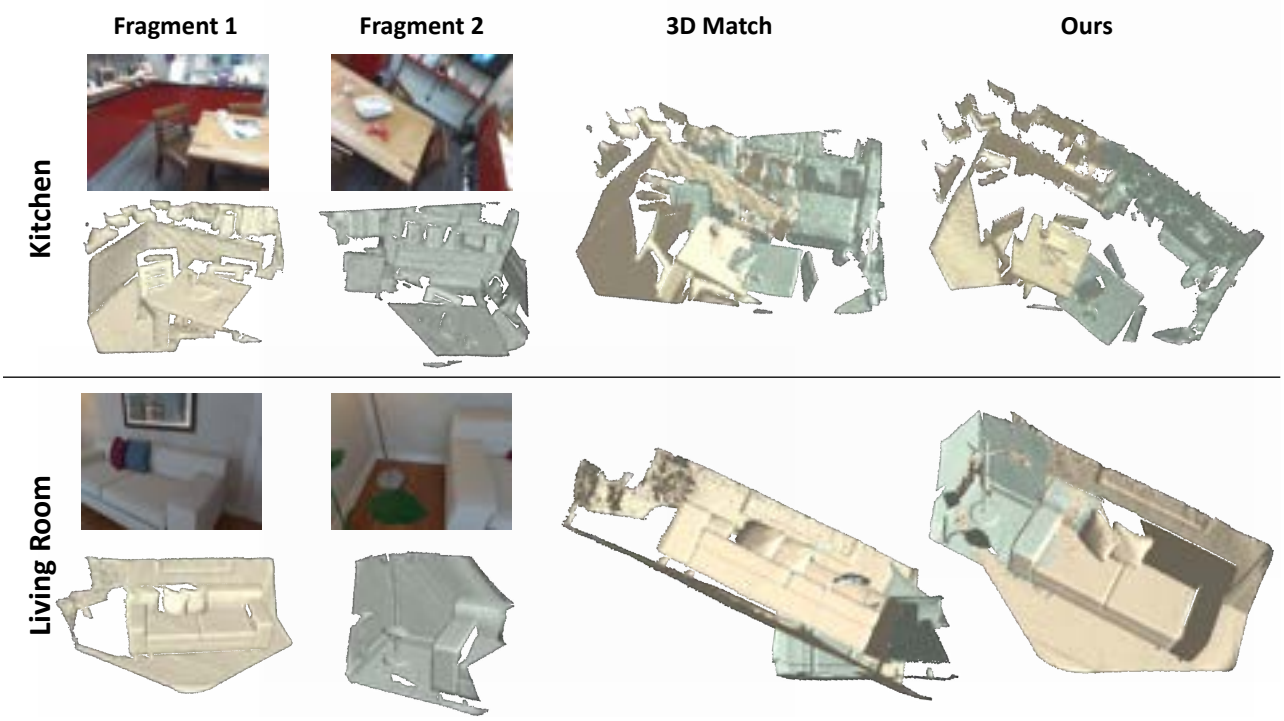

Figure 3: Examples of fragment alignment results using our learned 3D local descriptor. For comparison, we also provide the alignment results using the descriptors generated from the state-of-the-art 3DMatch [46]. From the visualization results, we can observe that our descriptor particularly performs better on the challenging cases that the two fragments only contain a small part in common. The color images of the fragments are used for better review only. No color image is used in our proposed method.

introduced by Choi et al. [8]. For each fragment pair (P_i, P_j), a computed transformation T_{ij} is compared to the ground-truth transformation T^*_{ij} only if $T_{ij}P_i$ overlaps certain percentage of P_j (e.g. 30%). T_{ij} is considered as a true positive if the RMSE of the ground-truth correspondences K^*_{ij} is below a threshold τ:

$$\frac{1}{|K^*_{ij}|} \sum_{(p^*, q^*) \in K^*_{ij}} ||T_{ij}p^* - q^*||^2 < \tau^2. \quad (6)$$

We used a threshold $\tau = 0.2$ in experiments on both synthetic fragment in the augmented ICL-NUIM [8] dataset and the real-world scan fragment datasets, including fragments from SUN3D [19, 44] and 7-Scenes [39]. Experimental settings and results will be discussed separately below.

Registration on synthetic fragments In this subtest, we train our model on the constructed training set mentioned in the subsection 4.1, and test it on the dataset provided by the 3DMatch authors, which contains some sampled volumetric point patches for the fragments in the augmented ICL-NUIM [8]. There are a total of 207 fragments in four scenes, including 57 fragments for *livingroom1*, 47 fragments for *livingroom2*, 53 fragments for *office1* and 50 fragments for *office2*.

Table 2 lists the average recalls and precisions of state-of-the-art algorithms on the augmented ICL-NUIM dataset. Our proposed method with adversarial enhancer can obtain a pretty high precision (28.3%) and recall (60.3%) . Specially, the precision is 3% higher than the most recent work 3DMatch. Moreover, the recall of our method is much higher than most of the compared methods, such as Spin-Images, FPFH. Though 3DMatch shows the highest recall among all the compared methods, it is computed with a fine-tuned model, which requires much more extra training time and training samples.

Registration on real-world fragments In addition to the synthetic fragments dataset, we also test our trained model on the dataset containing fragments from real-word sensor scans. The dataset consists of 60 fragments in the *redkitchen* scene of 7-scenes dataset, 120 fragments of *home* scene in SUN3D dataset, 103 fragments of *hotel* scene in SUN3D dataset, 66 fragments of *studyroom* scene in SUN3D dataset and 38 fragments of *lab* scene in SUN3D dataset. After extracting the local descriptors for the sampled point patches in each fragment, we compute the precisions and recalls given the transformations between each two fragments in the same scene.

The comparative results are provided in Table 3, where our proposed model without adversarial enhancer obtains 40.5% precision and 69.1% recall, and our model trained with adversarial enhancer performs even better with 72.0% precision and 42.9% recall. The

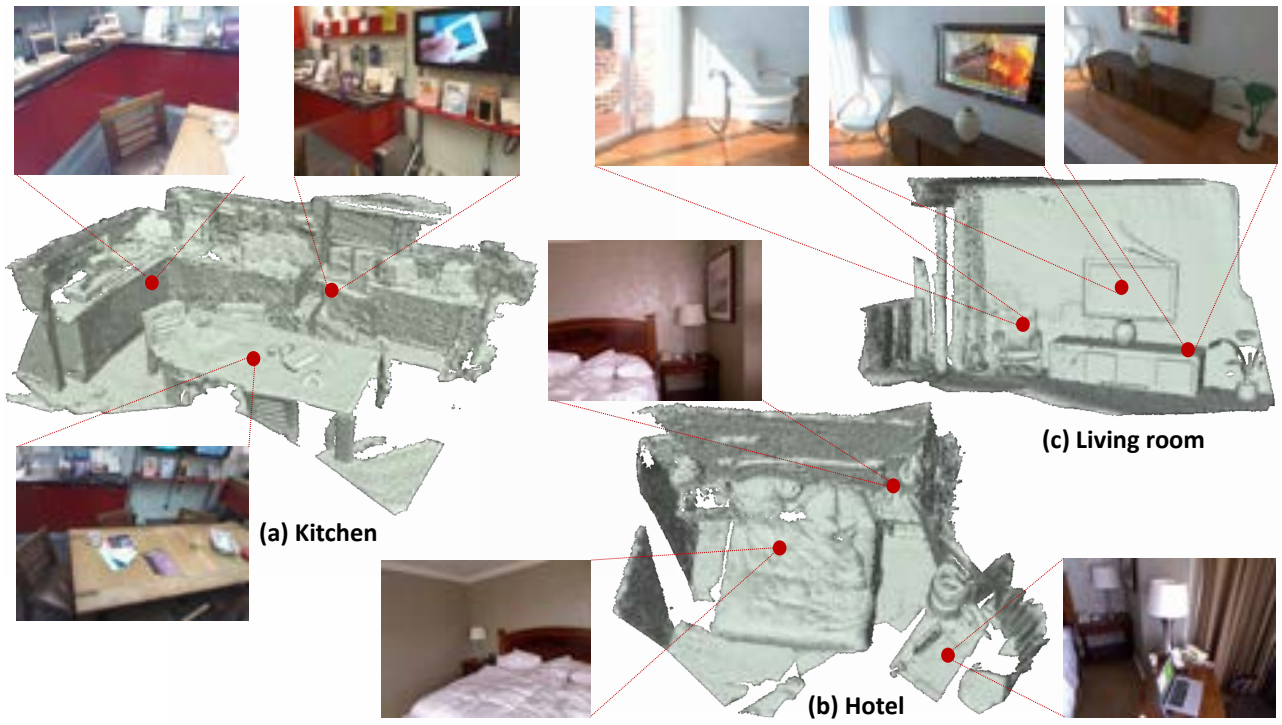

Figure 4: Examples of scenes that are reconstructed by simply combining multiple aligned fragments using our proposed descriptor. The color images of the scenes are displayed for better review only. Please note that no color image is needed in our proposed method.

large gap of the performance implies the great influence of introducing an adversarial enhancer when training the local descriptor generator. We also list the experimental results of other state-of-the-art methods in the table. As we can see from the table, our method performs the best among all the compared methods at both precision and recall measurements.

Qualitative analysis of fragment alignment Besides the quantitative evaluation, some examples of alignment results are visualized in Figure 3 for the intuitively qualitative alignment evaluation between two fragments within the same scene. In particular, we pick some alignments results using 3DMatch for comparison. All alignments are obtained with the transformation optimized by RANSAC algorithm. In some challenging cases (e.g. only a very small overlap between two fragments), 3DMatch could fail to align the fragments, but our descriptor is still able to handle such difficult cases properly. Also, we can reconstruct scenes only by simply combining multiple aligned fragments using our proposed descriptor. Figure 4 shows some of our scene reconstruction results, such as *kitchen* in 7-scenes dataset, *hotel* in SUN3D dataset and *living room* in ICL-NUIM dataset.

5 CONCLUSIONS

In this paper, we tackle the challenging 3D local matching problems by learning a robust local 3D descriptor with an adversarial enhancer. In order to enforce the local descriptor learned with the minimum distances between match pairs and the maximum distances between non-match pairs, we design a deep-siamese-network-based enhancer with a loss function opposite from the local descriptor generator. After training, given volumetric point patches, we extract the outputs from the learned local descriptor generator as their representations. The superior performance over state-of-the-art methods on both keypoint matching and geometry registration suggests that our proposed framework can learn a robust 3D local descriptor for volumetric point patches. To verify the effectiveness of our designed adversarial enhancer, we compare the keypoint matching and geometry registration performance of our framework with enhancer and without enhancer. The performance gap clearly indicates the training improvement of introducing the adversarial enhancer. Furthermore, the qualitative fragment alignment and scene reconstruction results demonstrate that our learned local descriptor can successfully match local point patches even in challenging cases that only a small common part exists between the fragments.

REFERENCES

[1] Mathieu Aubry, Ulrich Schlickewei, and Daniel Cremers. 2011. The wave kernel signature: A quantum mechanical approach to shape analysis. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 1626–1633.

[2] Serge Belongie, Jitendra Malik, and Jan Puzicha. 2002. Shape matching and object recognition using shape contexts. *IEEE transactions on pattern analysis and machine intelligence* 24, 4 (2002), 509–522.

[3] Davide Boscaini, Jonathan Masci, Simone Melzi, Michael M Bronstein, Umberto Castellani, and Pierre Vandergheynst. 2015. Learning class-specific descriptors for

deformable shapes using localized spectral convolutional networks. In *Computer Graphics Forum*, Vol. 34. Wiley Online Library, 13–23.
[4] Davide Boscaini, Jonathan Masci, Emanuele Rodolà, and Michael Bronstein. 2016. Learning shape correspondence with anisotropic convolutional neural networks. In *Advances in Neural Information Processing Systems*. 3189–3197.
[5] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. 2017. Unsupervised Pixel-Level Domain Adaptation With Generative Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3722–3731.
[6] Alexander M Bronstein, Michael M Bronstein, Leonidas J Guibas, and Maks Ovsjanikov. 2011. Shape google: Geometric words and expressions for invariant shape retrieval. *ACM Transactions on Graphics (TOG)* 30, 1 (2011), 1.
[7] Tsung-Han Chan, Kui Jia, Shenghua Gao, Jiwen Lu, Zinan Zeng, and Yi Ma. 2015. Pcanet: A simple deep learning baseline for image classification? *IEEE Transactions on Image Processing* 24, 12 (2015), 5017–5032.
[8] Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. 2015. Robust Reconstruction of Indoor Scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[9] Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol. 1. IEEE, 539–546.
[10] Dan Ciregan, Ueli Meier, and Jürgen Schmidhuber. 2012. Multi-column deep neural networks for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 3642–3649.
[11] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. 2017. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics (TOG)* 36, 3 (2017), 24.
[12] Yi Fang, Jin Xie, Guoxian Dai, Meng Wang, Fan Zhu, Tiantian Xu, and Edward Wong. 2015. 3d deep shape descriptor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2319–2328.
[13] Martin A Fischler and Robert C Bolles. 1987. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. In *Readings in computer vision*. Elsevier, 726–740.
[14] Ran Gal and Daniel Cohen-Or. 2006. Salient geometric features for partial shape matching and similarity. *ACM Transactions on Graphics (TOG)* 25, 1 (2006), 130–150.
[15] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2*. MIT Press, 2672–2680.
[16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).
[17] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. 2015. DRAW: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623* (2015).
[18] Kan Guo, Dongqing Zou, and Xiaowu Chen. 2015. 3d mesh labeling via deep convolutional neural networks. *ACM Transactions on Graphics (TOG)* 35, 1 (2015), 3.
[19] Maciej Halber and Thomas Funkhouser. 2017. Fine-To-Coarse Global Registration of RGB-D Scans. (2017).
[20] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. 2015. Deep transfer metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 325–333.
[21] Haibin Huang, Evangelos Kalogerakis, Siddhartha Chaudhuri, Duygu Ceylan, Vladimir G Kim, and Ersin Yumer. 2017. Learning local shape descriptors with view-based convolutional networks. *arXiv preprint arXiv:1706.04496* (2017).
[22] Andrew E. Johnson and Martial Hebert. 1999. Using spin images for efficient object recognition in cluttered 3D scenes. *IEEE Transactions on pattern analysis and machine intelligence* 21, 5 (1999), 433–449.
[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.
[24] Kevin Lai, Liefeng Bo, and Dieter Fox. 2014. Unsupervised feature learning for 3d scene labeling. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 3050–3057.
[25] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.
[26] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. 2017. Generative face completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[27] Ming-Yu Liu and Oncel Tuzel. 2016. Coupled generative adversarial networks. In *Advances in neural information processing systems*. 469–477.
[28] Jonathan Masci, Davide Boscaini, Michael Bronstein, and Pierre Vandergheynst. 2015. Geodesic convolutional neural networks on riemannian manifolds. In *Proceedings of the IEEE international conference on computer vision workshops*. 37–45.
[29] Nicolas Mellado, Dror Aiger, and Niloy J Mitra. 2014. Super 4pcs fast global pointcloud registration via smart indexing. In *Computer Graphics Forum*, Vol. 33. Wiley Online Library, 205–215.
[30] Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M Bronstein. 2017. Geometric Deep Learning on Graphs and Manifolds Using Mixture Model CNNs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5115–5124.
[31] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).
[32] Emanuele Rodolà, Samuel Rota Bulo, Thomas Windheuser, Matthias Vestner, and Daniel Cremers. 2014. Dense non-rigid shape correspondence using random forests. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4177–4184.
[33] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. 2009. Fast point feature histograms (FPFH) for 3D registration. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*. IEEE, 3212–3217.
[34] Radu Bogdan Rusu, Nico Blodow, Zoltan Csaba Marton, and Michael Beetz. 2008. Aligning point cloud views using persistent feature histograms. In *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*. IEEE, 3384–3391.
[35] Radu Bogdan Rusu, Zoltan Csaba Marton, Nico Blodow, and Michael Beetz. 2008. Learning informative point classes for the acquisition of object model maps. In *Control, Automation, Robotics and Vision, 2008. ICARCV 2008. 10th International Conference on*. IEEE, 643–650.
[36] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*. 2234–2242.
[37] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. 2013. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229* (2013).
[38] Lior Shapira, Shy Shalom, Ariel Shamir, Daniel Cohen-Or, and Hao Zhang. 2010. Contextual part analogies in 3D objects. *International Journal of Computer Vision* 89, 2-3 (2010), 309–326.
[39] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. 2013. Scene coordinate regression forests for camera relocalization in RGB-D images. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2930–2937.
[40] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. 2017. Semantic Scene Completion from a Single Depth Image. *Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition* (2017).
[41] Federico Tombari, Samuele Salti, and Luigi Di Stefano. 2010. Unique signatures of histograms for local surface description. In *European conference on computer vision*. Springer, 356–369.
[42] Xiaolong Wang, Abhinav Shrivastava, and Abhinav Gupta. 2017. A-Fast-RCNN: Hard Positive Generation via Adversary for Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2606–2615.
[43] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T Freeman, and Joshua B Tenenbaum. 2016. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in Neural Information Processing Systems*. 82–90.
[44] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. 2013. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 1625–1632.
[45] Li Yi, Hao Su, Xingwen Guo, and Leonidas Guibas. 2016. SyncSpecCNN: Synchronized Spectral CNN for 3D Shape Segmentation. *arXiv preprint arXiv:1612.00606* (2016).
[46] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 2017. 3DMatch: Learning Local Geometric Descriptors from RGB-D Reconstructions. In *CVPR*.
[47] Eugene Zhang, Konstantin Mischaikow, and Greg Turk. 2005. Feature-based surface parameterization and texture mapping. *ACM Transactions on Graphics (TOG)* 24, 1 (2005), 1–27.
[48] Jing Zhu, Jin Xie, and Yi Fang. 2018. Learning Adversarial 3D Model Generation With 2D Image Enhancer. In *the 32nd AAAI Conference on Artificial Intelligence (AAAI–18)*. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16064
[49] Jing Zhu, Fan Zhu, Edward K Wong, and Yi Fang. 2015. Learning pairwise neural network encoder for depth image-based 3d model retrieval. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 1227–1230.
[50] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2223–2232.