Metric Learning on Joint Embedding of 3D Scan and CAD Objects

Berna Kabadayi Angela Dai

Technical University of Munich

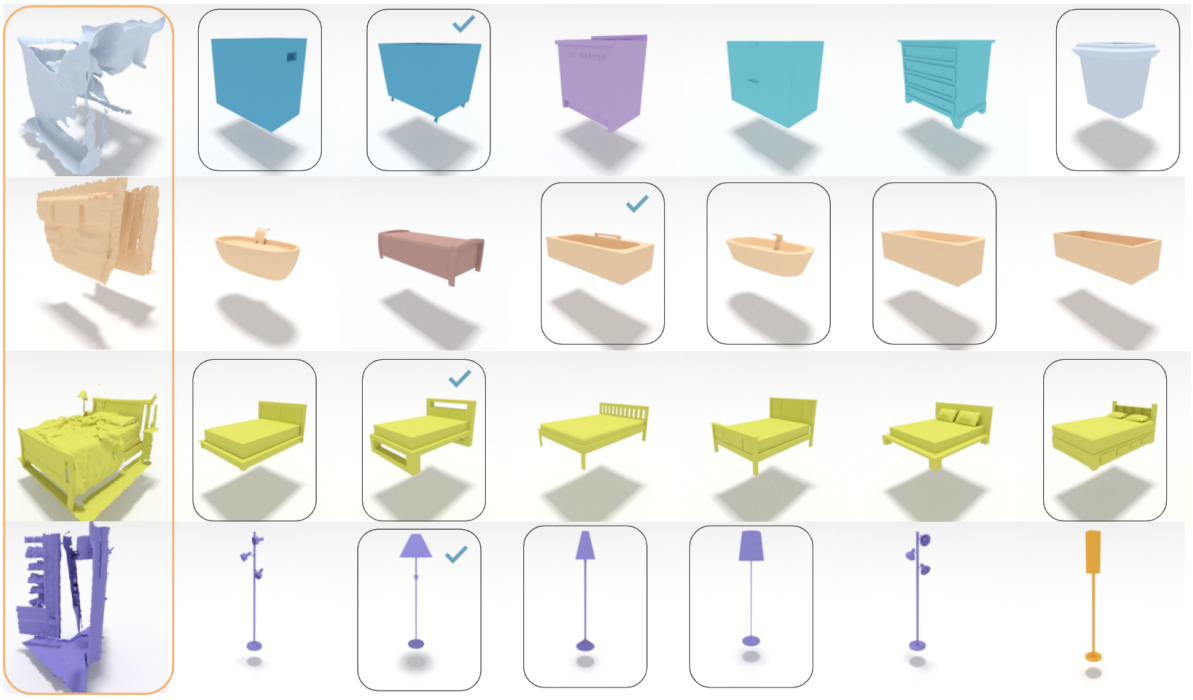

Figure 1: We improve the baseline [6], retrieves the the most similar CAD models from a partial scans, by 13% margin for the instance-level CAD model retrieval accuracy and by 18% margin for the Top-1 category-level CAD model retrieval accuracy by leveraging deep metric learning methods.

Abstract

With the help of the recent advances in 3D reconstruction algorithms, 3D scan geometry and scene understanding are getting popularity. Due to the noisy and incomplete nature of the 3D scan geometry, replacing scan models with the complete CAD models is important towards getting complete scenes. 3D scan geometry and CAD models contain interconnected information thus mapping between these two domains is crucial. In our work, based on the 3D CNN based approach proposed by Dahnert et al.[6], we leverage different metric learning and negative sampling methods to get a representative embedding space between these two domains. Our investigation by means of different metric learning such as contrastive loss and sampling approaches such as hard negative sampling outperforms our baseline [6] by 13% margin for the instance-level CAD model retrieval accuracy and by 18% margin for the Top-1 category-level CAD model retrieval accuracy.

1. Introduction

With the availability of the commodity sensors such as Microsoft Kinect, Intel RealSense which provides pixel-by-pixel depth images besides RGB frames, 3D reconstruction of the scenes has been advanced for the several years within computer vision and graphics community [17, 13, 30].

3D reconstruction of the static and dynamic scenes helps us to understand the underlying representation of the 3D

geometry by leveraging object detection, semantic segmentation tasks [12] on it. However, the 3D geometry that is reconstructed by the standard RGB-D reconstruction pipeline [13, 17] is often not-complete, noisy and still way more far from compact representation of the scene.

Therefore, to tackle missing and noisy geometry problem, several algorithms are proposed in the direction of either completing the missing parts of the 3D geometry in a supervised [9] or self-supervised manner [8] or replacing the noisy and incomplete real-world scan objects with the CAD models [6, 23, 19] with the help of the increasing availability of public CAD model datasets [1]. Both real-world scan objects and 3D CAD models are commonly used for semantic understanding of environment and contain compromising information in a way that CAD models are compact, clean and simple, whereas real-world scan objects are more complex, noisy and incomplete. Complete and clean representation of the 3D environment by means of CAD models helps us to operate different tasks such as semantic segmentation, object detection, which will be useful for many applications such as AR-VR. In that sense, finding a representative feature mapping in between the scan objects and CAD models is necessary for scan-CAD model retrieval problem.

Benefiting from the of publicly available large scale datasets such as ScanNet [7] and ShapeNet [1] which provide large amount of scan objects and CAD models, respectively, several algorithms are proposed for finding a joint embedding space of the scan objects and CAD models. [6] Most of the approaches focus on the category level scan-CAD retrieval [6, 23, 19]. Category level retrieval is that if the class of retrieved CAD model is the same with the class of the query scan object, then this is considered as a correct retrieval. In spite of the fact that class level retrieval provides us a mapping between two domains, capturing intra-class mappings would offer more meaningful semantic information. Recently, Dahnert et al. [6] proposed a new 3D CNN based approach to discover intra-class level similarities.

In this research work, by using Joint Embedding of 3D Scan and CAD Objects paper [6] as a baseline, we asked two main questions for instance-level scan-CAD model retrieval. First, can we leverage from 3D deep metric learning methods besides standart triplet loss? Second, can we benefit from sampling methods during training to leverage intra-class similarities? In order to answer the first question, we investigate the effect of commonly used deep metric learning methods such as contrastive loss [10], quadruplet loss [3] on the scan-CAD similarity benchmark[6]. For the second question, we focus on the online hard negative mining during training and compare the performance with a combination of embedding losses. Our method outperforms the state of the art baseline with 15% margin for instance-level retrieval accuracy and by 18% margin for Top-1 category-level retrieval accuracy.

In summation, we make the following contributions in this research work on top of our baseline paper:

- We investigate the effect of different metric learning methods for 3D representations and find out that contrastive loss outperforms the triplet loss for scan-CAD model retrieval problem.
- We benefit from different sampling methods during training and show that online negative mining within the same class helps the reveal intra-class similarities.

2. Related Work

CAD-Model Retrieval With the availability of the large scale richly annotated scan and CAD model datasets such as ScanNet [7] for scan models, ShapeNet [1] for CAD models, CAD model retrieval methods can be considered as an expressive way of representing the image or the scene in a complete fashion. There are several directions in 3D CAD model retrieval problem.

The first approach is that researchers tackle the problem as an RGB-D to CAD model retrieval where given an RGB-D object or a scene, the most similar CAD models should be retrieved by the method. SHREC 17 [23] and SHREC 18 [19], RGB-D to CAD retrieval challenges, focus on the class level retrieval. These challenges helps us in many ways such as scene modelling or better AR/VR applications. It is also important to capture the instance level differences as well as the class level differences. Instance level CAD model retrieval method aims to investigate the differences in the same class. (i.e. retrieving the correct instance of the class rather than just predicting which class object belongs to). Dahnert et al. [6] addresses that direction by providing a new approach to capture the instance level 3D CAD model retrieval.

Besides considering the problem as an RGB-D to CAD model retrieval, the second direction is the image based 3D CAD model retrieval. Given the RGB image captured in the real world, the methods aim to estimate the most similar and relevant CAD models [15]. In addition, Mask2CAD [14] formulates the problem as an image-shape embedding learning. Yuan et al. [29] also proposes a method which estimates 6D object pose including its orientation and location.

In the scope of this research work, we focus on the RGB-D to CAD model retrieval problem by using Dahnert et al. [6] work as a baseline.

Deep Metric Learning Deep metric learning tackles the problem of mapping the high dimensional data into the meaningful embedding space so that these embeddings can

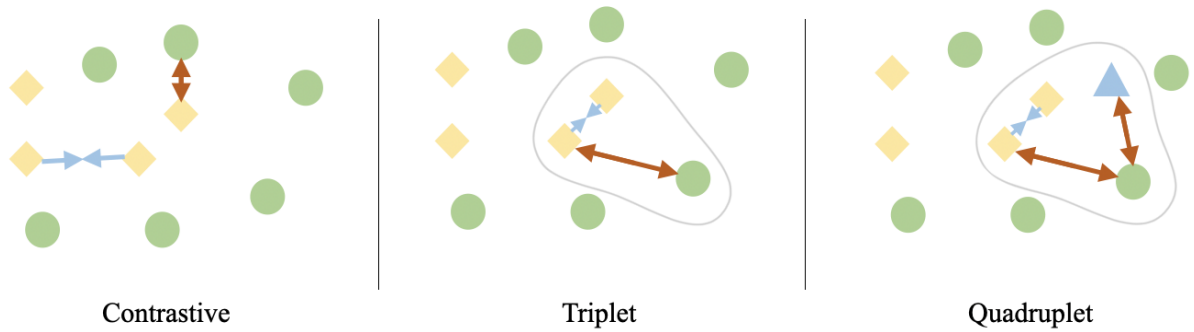

Figure 1. Random negative sampling scenario for Contrastive [10], Triplet[11], Quadruplet[3] loss formulation where green, yellow and blue objects represent different classes. The visualization is prepared with the help of the Choy et al. [5].

be used for different tasks such as face recognition, object retrieval, etc.

The researches tackle the metric learning either as a classification based losses or embedding based losses. Classification based losses such as normalized softmax loss formulates the weight matrices into the class logits. However, embedding based losses such as contrastive loss try to learn the relations available in the batch during training. Therefore, how batches are sampled is also a very important concern in the embedding losses as well to be able to learn the semantic differences in between the data.

More than a decade ago, Hadsell et al. [10] proposed a dimensionality reduction method, which is known as a standart constrastive loss, based on the considering neighborhood relationship of the data. The idea of the contrastive loss is pulling the similar samples into each other while pushing the dissimilar samples from each other.

Another type of commonly used embedding loss is triplet loss which is introduced in the FaceNet [11]. It aims to learn the representations by comparing the distances between positive and and negative sample given an anchor.

More recently, MoCo [4], SimCLR [2] and PIRL [16] propose methods to learn the visual representations in a self-supervised way.

3. Method Overview

Our hourglass fashioned network [18] learns the joint embedding space between real-world scan objects and CAD models with the help of the fully 3D convolutional networks. The network consists of two stacked hourglass followed by an final encoder to find latent space between scan objects and CAD models. The first and second hourglass encoder-decoders help us to obtain more CAD-like representations of the real-world scan objects before mapping them with the CAD models to the joint embedding space.

The first hourglass which consists of an encoder-decoder learns to segment foreground object from background with the help of the provided scan object mask. The second hourglass which also consists of an encoder-decoder takes partial scan object that is obtained from the first hourglass and a reference CAD model as an input, then learns to complete the partial scan object. The first and second hourglass help to obtain more CAD-like representation. Then, the last encoder takes scan object whose background is removed and partialness is completed, a positive CAD model and a randomly generated negative CAD model as an input, then produces feature vectors of the 3D models to find the joint embedding space between real-world scan object and CAD models by formulating triplet loss between scan-object, positive CAD model and negative CAD model [6]. The scan objects and CAD models are all represented as a 32^3 binary occupancy grids.

We formulate different loss functions on 32^3 dimensional scan object, positive CAD model that corresponds to scan object and negative CAD model which is sampled in several ways. In the results section, we demonstrate the effect of loss functions and sampling strategies in detail.

This end-to-end 3D CNN based method [6] learns the shared embedding space of real-world scan objects and CAD models and helps us to reveal intra-class similarities between CAD models in the same class for many applications such as CAD model retrieval.

4. Metric Learning

In this section, we briefly mention about standard loss functions such as contrastive loss and triplet loss for metric learning and commonly used negative mining techniques.

4.1. Loss Functions

Contrastive loss The standard contrastive loss [10] takes a pair of embedding vectors and a flag stating whether they are similar samples or not. If they are similar samples contrastive loss tries to minimize the difference between them. If they are dissimilar, then constrastive loss tries to max-

imize the distance up to the some margin. The proposed margin based contrastive loss:

$$\mathcal{L}(x_1, x_2, y) = y * d(x_1, x_2) + (1 - y) * \max\{0, m - d(x_1, x_2)\} \quad (1)$$

where $d(x_1, x_2)$ is a distance metric which is often L2 distance between feature vectors, Y is a label of the similarity and m is margin for negative samples. Choy et al. [5] uses margin for also positive pairs to prevent overfitting.

Triplet Loss The standard triplet loss[26] takes an anchor, positive and negative samples. The triplet loss tries to learn a metric where positive sample is closer to the anchor than the negative sample which is dissimilar to anchor up to the some non-negative margin. The proposed margin based triplet loss for three feature vectors x_1, x_2 and x_3:

$$\mathcal{L}(x_1, x_2, x_3) = \max\{0, d(x_1, x_2) - d(x_1, x_3) + m\}$$

where $d(x_1, x_2)$ is a distance between anchor and positive sample, $d(x_1, x_3)$ is a distance between anchor and negative sample and m is a non negative margin.

The difference between standard contrastive and triplet loss is that the margin in contrastive loss is based on the exact distances between two embeddings. However, the triplet loss considers the relative distance between pairs.

Quadruplet Loss The quadruplet loss [3] can be considered as an extension of the standard triplet loss. It is claimed that quadruplet loss pushes away the negative pairs from positive pairs and enables better generalization by introducing a new constraint on the negative samples. The proposed quadruplet loss:

$$\mathcal{L}(x_1, x_2, x_3, x_4) = d(x_1, x_2) - d(x_1, x_3) + m_1 + d(x_1, x_2) - d(x_3, x_4) + m_2 \quad (2)$$

where $d(x_1, x_2)$ is a distance between anchor and positive sample, $d(x_1, x_3)$ is a distance between anchor and negative samples, $d(x_3, x_4)$ is a distance between two negative samples coming from different classes and m_1 and m_2 is a non negative margins.

In the context of scan-CAD model retrieval task, we try above-mentioned loss functions as shown in the Figure 1 and their variations.

4.2. Negative Sampling Mining

In metric learning, there are three types of negative samples which are easy, semi-hard and hard negatives. In this section, we briefly explain what they are for triplet loss and focus on the hard negative sampling in our work like most of the recent approaches [5].

Easy Triplets The triplet loss becomes easily 0. Because the negative sample is located relatively far away from the anchor considering positive one as shown in the Figure 2.

$$d(a, p) + m < d(a, n)$$

Semi-Hard Triplets In these triplets, the positive sample is closer than negative but we still have positive triplet loss.

$$d(a, p) < d(a, n) < d(a, p) + m$$

Hard Triplets The negative sample is closer to the anchor than positive.

$$d(a, n) < d(a, p)$$

In the scope of the scan-CAD model retrieval, we leverage different sampling methods considering the above-mentioned negatives, since the way we sample during the training helps us to learn the geometric representations better. For this work, we focus on the online mining techniques where samples are randomly chosen during training instead of the offline mining techniques such as memory bank approach [27].

The first approach shown in the Figure 2 which is also used in the baseline [6] is that we remove the all instances within the same class for a given query in a batch, then randomly select one CAD model from the remained ones. Instead of using triplet loss we use contrastive loss. It is called random negative sampling. The result of random negative sampling for the contrastive loss scenario is shown in Table 2. This scenario handles the easy negatives.

One drawback of the first method is that although this random negative sampling helps us to learn the inter-class representations, sampling randomly with the CAD model from different class does not allow to capture 3D representations sufficiently in the same class. To tackle hard negatives the way we sample is that we first remove the instance of the CAD model from the batch, then randomly sample from the same class then assign a relatively very small margin to be able to leverage the features in the same class. This sampling is shown in Figure 3.

Latter-mentioned sampling does not guarantee that after removing the CAD instance from the batch we will have another CAD sample from the same class for sampling. Therefore, we propose an adaptive margin sampling approach for the negative samples. In the minibatch we construct, we first remove the corresponding CAD model of the scan model then check the remaining ones from the same class. If there is still CAD samples from the same class then we sample randomly from the same class with a relatively small margin, otherwise we randomly sample from different class and assign a large margin. This is shown in the Figure 4.

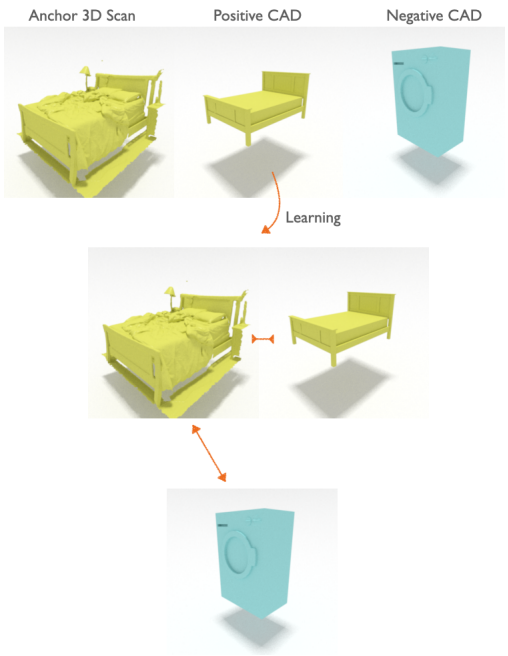

Figure 2. Traditional triplet loss [24] in a random negative sampling scenario. It reduces the difference between the similar objects (i.e 3D scan-positive CAD) and increase the difference between different objects (i.e 3D scan-negative CAD).

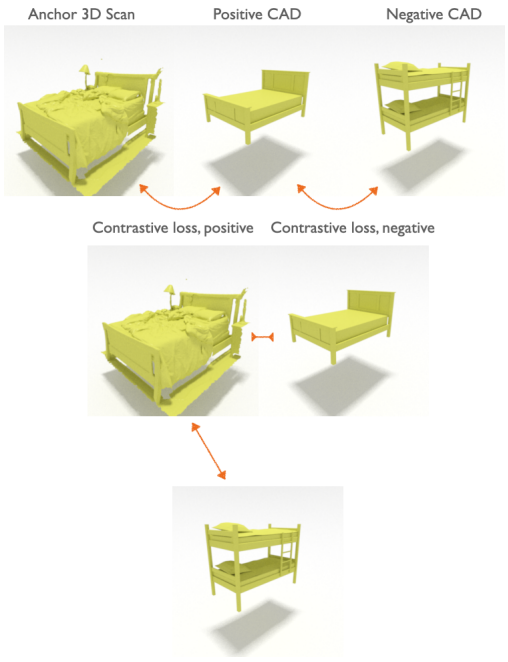

Figure 3. Hard negative sampling within the same class.

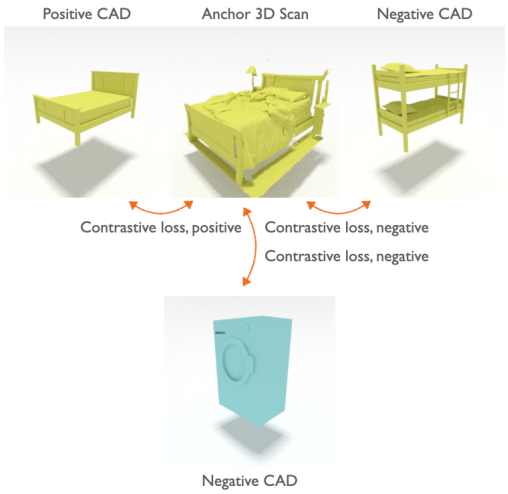

Figure 4. Adaptive margin approach: Hard negative sampling within the same class if available in the batch and set a small margin, otherwise sampling from the different class and set a large margin.

5. Network Training

Dataset For training we use the same setup with the baseline paper [6]. There are 14123 samples for training, validation and test in total. Each sample is constructed from a real-world scan object and segmentation mask from ScanNet[7], corresponding positive CAD model from ShapeNet [1]. The training, validation and tests splits consist of 9571, 1398 and 3154 scan-CAD pair samples, respectively.

For testing, we use the public Scan-CAD Object Similarity Dataset[6]. It offers 5102 annotated scan-CAD pairs. For a query scan there are 6 CAD models which are proposed for annotation to user and 3 of them marked as most similar ones. There are 3207, 814 and 1081 annotated samples available for training, validation, test splits, respectively.Within Scan-CAD Object Similarity Dataset[6], 2554 out of 3207 training samples are unique scans. 578 of the 814 validation samples and 847 out of the 1081 test samples offer unique scans. The sample from Scan-CAD Object Similarity Dataset is shown in the Figure 5.

We also benefit from the training and validation parts of the scan-CAD Object Similarity Dataset[6] for finetuning purposes.

Network Architecture The baseline network architecture consists encoders followed by decoders. The first encoder-decoder component is responsible for foreground-background segmentation. The second encoder-decoder is for scan completion. The last encoder is Siamese network for 3 input. For funetuning we modified the last encoder in a way that we can forward 6 CAD models for every scan

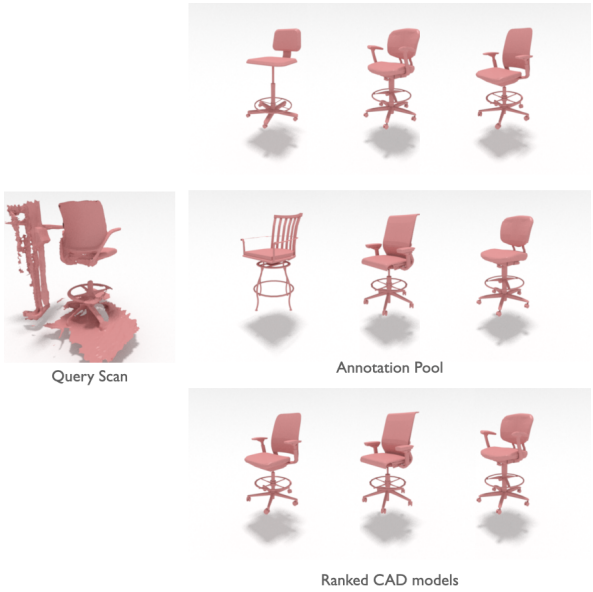

Figure 5. Sample from Scan-CAD Object Similarity dataset. For every scan, there are 6 CAD models in the pool, 3 CAD models ranked as the most similar ones.

model, which are available in the Scan-CAD Similarity Benchmark Dataset.

Optimization We use Adam optimzer with a batch size 128 for end to end training. We use the learning rate $1e{-}3$ with a step-wise 0.1 learning rate decay ratio. We train our model around 1 day for 800 to 1000 epochs on a single Nvidia GTX 1080Ti. For finetuning, we decrease the batch size from 128 to 64 for memory limitations and train around 5 hours on the Scan-CAD Similarity Benchmark Dataset.

6. Results

The network takes RGB-D scan as an input, and learns to segment background and foreground from each other then completes the missing parts of the scan with provided appropriate CAD model. Then, the network tries to learn joint embedding space between scan and CAD geometry by formulating triplet loss in the baseline [6]. What we propose instead of a standard triplet loss formulation directly, we explore several directions and explain the results in this section.

Effect of Contrastive Loss We first run experiments on the contrastive loss with positive and negative margins instead of a standard triplet loss which is used in the baseline model [6]. We show that contrastive loss with a negative and positive margin outperforms the triplet loss for revealing geometric features. The best result we get with contrastive loss formulation is that we achieve the 0.53 instance retrievel accuracy. where we use 1.25 negative margin and random negative sampling. The detailed comparison on the contrastive loss with different margins can be found the Table 2.

Negative Sampling for Contrastive Loss Secondly, we experiment different sampling techniques for contrastive loss formulation. In the baseline we use, the negative CAD models are randomly sampled from different classes within the same batch. In addition to that, we also sample negatives from the same class. The hard-negative sampling performed near random sampling in contrastive loss setting but still outperforms the standart triplet loss by 7% margin. The detailed comparison of the can be found in Table 2

Effect of Quadruplet Loss Quadruplet Loss near behaved triplet loss baseline. It is likely that it enforces regularization between negative samples in which we leverage less from in the scope of the scan-CAD object similarities.

Finetuning on Scan-CAD Last, we benefit from the ranked scan-CAD object similarities dataset which is provided by our baseline paper [6]. We finetune one of the our best models that we evaluate on the similarity dataset with the contrastive loss. We change the last encoder architecture from 3 CAD sample to 6 CAD samples and applied easy positive mining strategy [28]. The reason why we fine-tune is that the standart test setup is way more challenging than training scenario. In training, the network learns to differentiate positive and negative CAD samples, whereas in test the networks predicts the most similar CAD models provided 6 CAD models within the same class. We achieve 0.58 instance retrieval accuracy, which is state of the art performance on scan-CAD Object Similarity Benchmark. Fine-grained evaluation scores for retrieval accuracy and ranking quality is shown in Table 4.

6.1. Quantitative Results - Metrics

Top1 Category-based Retrieval Accuracy Category-based retrieval accuracy is that if the retrieved class is the same with the query scan class then this retrieval is marked as true positive as shown in 6. Our method outperforms the baseline [6] and compared methods by %18 for the Top-1 Category-based retrieval accuracy as shown in Table 1.

Instance Retrieval Accuracy Instance retrieval accuracy, proposed by Dahnert et al. [6], is improved version of category-level retrieval accuracy. Instance level retrieval accuracy is that for a given scan query, annotated 6 CAD samples, ranked 3 CAD samples, if the methods retrieve one of the CAD models from ranked list, then it is considered as true positive. The example sample

Figure 6. Sample from Scan-CAD Object Similarity dataset. For every scan, there are 6 CAD models in the pool, 3 CAD models ranked as the most similar ones.

Method	Top-1 Retrieval Accuracy
FPFH [22]	0.14
SHOT[25]	0.07
PointNet [20]	0.49
3DCNN [21]	0.57
JointEmbedding [6]	0.68
Our method	**0.86**

Table 1. Top 1 category-based CAD model retrieval accuracy comparison with the baseline we have [6] and other methods compared with. By using contrastive loss and negative sampling strategy, we improve the Top-1 category-based retrieval accuracy by %18. Previous experiments are taken from Joint Embedding of 3D Scan and CAD Objects [6].

from Scan-CAD Object Similarity dataset is shown in Figure 5. In Table 2 and Table 5, instance retrieval accuracy results are shown. Visual results are shown in Figure 1.

Ranking Quality Ranking quality, proposed by Dahnert et al. [6] is that top n $(n \leq 3)$ predicted CAD models are compared with the ranked ones, evaluate number of commons and divide by n. The comparison of the ranking quality is shown in the Table 3.

7. Limitations

Although our deep metric learning and sampling approaches outperform the baseline we use [6], there are several limitations. For instance, we only consider the geometric information, therefore, combining scan-CAD model retrieval problem with the image-CAD model retrieval task would potentially be another direction.

8. Conclusion

By having a baseline on the 3D CNN based approach to find mapping between scan-CAD model domains, we leverage different metric learning and sampling algorithms to learn the representation of this domains better. We show that contrastive loss and hard negative sampling approaches improve not only category level retrieval evaluation but also instance level retrieval evaluation.

9. Acknowledgements

All this mentioned work is implemented as a part of the interdisiplinary research project at TUM. We would like to thank Visual Computing Lab, especially Matthias Nießner for computing resources and Manuel Dahnert for providing the baseline code.

References

[1] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. Shapenet: An information-rich 3d model repository, 2015. 2, 5

[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations, 2020. 3

[3] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification, 2017. 2, 3, 4

[4] X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[5] C. Choy, J. Park, and V. Koltun. Fully convolutional geometric features. In *ICCV*, 2019. 3, 4

[6] M. Dahnert, A. Dai, L. Guibas, and M. Niessner. Joint embedding of 3d scan and cad objects. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. 1, 2, 3, 4, 5, 6, 7, 9, 10

[7] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 2, 5

[8] A. Dai, C. Diller, and M. Nießner. Sg-nn: Sparse generative neural networks for self-supervised scene completion of rgb-d scans. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2020. 2

[9] A. Dai, D. Ritchie, M. Bokeloh, S. Reed, J. Sturm, and M. Nießner. Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2018. 2

[10] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, page 17351742, USA, 2006. IEEE Computer Society. 2, 3

[11] E. Hoffer and N. Ailon. Deep metric learning using triplet network. *Lecture Notes in Computer Science*, page 8492, 2015. 3

[12] J. Hou, A. Dai, and M. Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2019. 2

[13] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *In Proc. UIST*, pages 559–568, 2011. 1, 2

[14] W. Kuo, A. Angelova, T.-Y. Lin, and A. Dai. Mask2cad: 3d shape prediction by learning to segment and retrieve, 2020. 2

[15] W. Li, A. Liu, W. Nie, D. Song, Y. Li, W. Wang, S. Xiang, H. Zhou, N.-M. Bui, Y. Cen, Z. Chen, H.-H. Chung-Nguyen, G.-H. Diep, T.-L. Do, E. L. Doubrovski, A.-D. Duong, J. Geraedts, H. Guo, T.-H. Hoang, Y. Li, X. Liu, Z. Liu, D. Luu, Yun-sheng, Ma., V. Nguyen, J. Nie, T. Ren, M.-K. Tran, S.-T. Tran-Nguyen, M. Tran, T.-A. Vu-Le, C. Wang, S. Wang, G. Wu, C. Yang, M. Yuan, H. Zhai, A. Zhang, F. Zhang, and S. Zhao. Shrec 2019-monocular image based 3 d model retrieval. 2019. 2

[16] I. Misra and L. van der Maaten. Self-supervised learning of pretext-invariant representations. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2020. 3

[17] R. A. Newcombe, S. Izadi, O. Hilliges, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE ISMAR*. IEEE, October 2011. 1, 2

[18] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. *CoRR*, abs/1603.06937, 2016. 3

[19] Q.-H. Pham, M.-K. Tran, W. Li, S. Xiang, H. Zhou, W. Nie, A. Liu, Y. Su, M.-T. Tran, N.-M. Bui, et al. Shrec18: Rgb-d object-to-cad retrieval. *Proc. 3DOR*, 2, 2018. 2

[20] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CoRR*, abs/1612.00593, 2016. 7

[21] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. J. Guibas. Volumetric and multi-view cnns for object classification on 3d data. *CoRR*, abs/1604.03265, 2016. 7

[22] R. B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (fpfh) for 3d registration. ICRA'09, page 18481853. IEEE Press, 2009. 7

[23] M. Savva, F. Yu, H. Su, A. Kanezaki, T. Furuya, R. Ohbuchi, Z. Zhou, R. Yu, S. Bai, X. Bai, M. Aono, A. Tatsuma, S. Thermos, A. Axenopoulos, G. T. Papadopoulos, P. Daras, X. Deng, Z. Lian, B. Li, H. Johan, Y. Lu, and S. Mk. Large-Scale 3D Shape Retrieval from ShapeNet Core55. In I. Pratikakis, F. Dupont, and M. Ovsjanikov, editors, *Eurographics Workshop on 3D Object Retrieval*. The Eurographics Association, 2017. 2

[24] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015. 5

[25] F. Tombari, S. Salti, and L. Di Stefano. Unique signatures of histograms for local surface description. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Computer Vision – ECCV 2010*, pages 356–369, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. 7

[26] D. P. Vassileios Balntas, Edgar Riba and K. Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In E. R. H. Richard C. Wilson and W. A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 119.1–119.11. BMVA Press, September 2016. 4

[27] Z. Wu, Y. Xiong, S. Yu, and D. Lin. Unsupervised feature learning via non-parametric instance-level discrimination, 2018. 4

[28] H. Xuan, A. Stylianou, and R. Pless. Improved embeddings with easy positive triplet mining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 6

[29] H. Yuan, R. C. Veltkamp, G. Albanis, N. Zioulis, D. Zarpalas, and P. Daras. SHREC 2020 Track: 6D Object Pose Estimation. In T. Schreck, T. Theoharis, I. Pratikakis, M. Spagnuolo, and R. C. Veltkamp, editors, *Eurographics Workshop on 3D Object Retrieval*. The Eurographics Association, 2020. 2

[30] M. Zollhöfer, P. Stotko, A. Görlitz, C. Theobalt, M. Nießner, R. Klein, and A. Kolb. State of the art on 3d reconstruction with rgb-d cameras. *Computer Graphics Forum (EG STAR)*, 37(2):625–652, May 2018. 1

Method	Margin (neg,pos)	trash bin	bathtub	bed	bookshelf	cabinet	chair	display	file	sofa	table	other	**class avg**	**inst avg**
Baseline	$m_{triplet} = 0.2$	0.50	0.60	0.42	0.19	0.26	**0.55**	0.45	0.25	0.33	0.32	0.43	0.39	0.43
Random neg sampling	$m_n = 1.0$	0.55	0.49	0.44	0.55	0.32	0.48	0.46	0.41	0.54	0.48	0.40	0.46	0.47
Random neg sampling	$m_n = 1.25$	**0.81**	0.48	0.49	0.57	0.29	0.54	0.51	0.38	0.49	0.46	0.44	0.50	**0.53**
Random neg sampling	$m_n = 1.50$	0.60	0.61	0.32	**0.74**	0.38	0.46	0.38	0.39	0.41	0.46	0.42	0.47	0.48
Random neg sampling	$m_n = 1.25, m_p = 1.25$	0.55	0.47	0.47	0.42	0.34	0.49	0.57	0.36	0.51	0.44	0.33	0.45	0.46
Random neg sampling	$m_n = 1.25, m_p = 0.1$	0.70	0.50	0.51	0.52	0.40	0.45	**0.56**	0.41	0.51	0.48	0.40	0.49	0.49
Same and diff. class	$m_n = 1.25, 0.1, m_p = 0$	0.42	0.58	0.42	0.46	0.37	**0.55**	0.39	**0.47**	0.49	**0.55**	**0.52**	0.47	0.50
Same and diff. class	$m_n = 1.25, 0.2, m_p = 0$	0.44	0.59	0.51	0.51	0.35	0.54	0.41	0.42	0.48	**0.55**	0.50	0.48	0.50
Same and diff. class	$m_n = 1.25, 0.3, m_p = 0$	0.40	**0.62**	**0.54**	0.50	0.39	0.54	0.47	0.43	**0.55**	0.52	0.48	0.49	0.50
Random sampling&reg	$m_n = 1.0 m_p = 0$	0.75	0.38	0.37	0.40	**0.47**	0.45	0.49	0.38	0.52	0.48	0.43	0.46	0.49
Random sampling&reg	$m_n = 1.25 m_p = 0$	0.76	0.53	0.49	0.64	0.34	0.52	**0.56**	0.41	0.53	0.47	0.40	**0.51**	0.52

Table 2. Top-1 Retrieval Accuracy comparison of baseline paper [6] and our contrastive loss setting per class on Scan-CAD Object Similarity benchmark public dataset with different margins. The explanation of the methods as follows. **Random neg sampling:** Random negative sampling from different class after removing query object CAD from the batch. **Same and diff class:** After removing query object CAD from the batch, if there is at least one sample CAD within the same class then sample with it assign small negative margin, if not then sample with random negative from different class. **Random sampling&reg:** Random negative sampling from different class after removing query object CAD from the batch. In addition add constraint like positive CAD and negative sampled CAD also should be far away from each other up to some margin.

Method	Margin (neg,pos)	trash bin	bathtub	bed	bookshelf	cabinet	chair	display	file	sofa	table	other	class avg	inst avg
Baseline	$m_{triplet} = 0.2$	0.29	0.24	0.19	0.08	0.12	0.19	0.14	0.19	0.15	0.10	0.09	0.16	0.16
Random neg sampling	$m_n = 1.0$	0.22	0.21	0.19	0.24	0.14	0.17	0.15	0.13	0.19	0.18	0.17	0.18	0.18
Random neg sampling	$m_n = 1.25$	0.37	0.20	0.17	0.25	0.14	0.19	0.18	0.14	0.19	0.17	0.19	**0.20**	0.20
Random neg sampling	$m_n = 1.50$	0.23	**0.28**	0.17	**0.30**	0.15	0.18	0.17	0.17	0.14	0.16	0.18	0.19	0.18
Random neg sampling	$m_n = 1.25, m_p = 1.25$	0.28	0.16	0.17	0.19	0.15	0.18	0.17	0.15	0.19	0.16	0.13	0.17	0.18
Random neg sampling	$m_n = 1.25, m_p = 0.1$	0.37	0.17	0.22	0.22	0.15	0.17	0.15	0.16	0.16	0.16	0.20	0.19	0.19
Same and diff. class	$m_n = 1.25, 0.1, m_p = 0$	0.20	0.31	0.19	0.20	0.16	0.19	0.17	**0.22**	**0.21**	0.18	0.19	**0.20**	0.19
Same and diff. class	$m_n = 1.25, 0.2, m_p = 0$	0.18	0.23	0.18	0.20	0.15	**0.21**	**0.22**	0.19	0.18	**0.20**	**0.22**	**0.20**	**0.20**
Same and diff. class	$m_n = 1.25, 0.3, m_p = 0$	0.18	0.24	**0.21**	0.21	0.15	0.20	0.19	0.21	0.16	0.18	0.18	0.19	0.19
Random sampling&reg	$m_n = 1.0 m_p = 0$	**0.38**	0.18	0.20	0.18	**0.17**	0.18	0.13	0.20	0.16	0.18	0.17	0.19	0.19
Random sampling&reg	$m_n = 1.25 m_p = 0$	0.34	0.23	0.16	0.24	0.15	0.19	0.19	0.15	0.16	0.17	0.17	0.19	**0.20**

Table 3. Top 1 ranking quality comparison of baseline paper [6] and our contrastive loss setting per class on Scan-CAD Object Similarity benchmark public dataset. The explanation of the methods can be found in the Table 2.

Contrastive Loss Formulation with 6 CAD models	Margin	Train		Validation		Test		Instance Avg
		All	Filtered	All	Filtered	All	Filtered	
Pushing the negatives, pulling the positives	$m_n = 0.5$	0.46	0.50	0.46	0.55	0.44	0.47	0.45
Pushing the negatives, pulling the positives	$m_n = 1.2$	0.49	0.53	0.49	0.58	0.50	0.55	0.49
Only pushing the negatives	$m_n = 1.4$	0.44	0.49	0.46	0.57	0.41	0.47	0.43
Pushing the negatives, pulling the positives	$m_n = 2.0$	0.56	0.61	0.54	0.61	0.51	0.55	0.54
Pushing the negatives, pulling the positives	$m_n = 2.5$	**0.60**	**0.63**	**0.55**	0.63	**0.53**	**0.56**	**0.58**
Pushing the negatives, pulling the positives	$m_n = 2.75$	0.56	0.63	0.54	**0.64**	**0.53**	**0.56**	0.55
Pushing the negatives, pulling the positives	$m_n = 3.0$	0.53	0.59	0.51	0.63	0.51	0.54	0.52

Table 4. Finetuning results for Top-1 Retrieval Accuracy on the Scan-CAD Object Similarity Benchmark [6]. For every sample, there are 6 CAD models which are used for annotation. 3 of 6 CAD models are annotated as the most similar ones. However, since these ranked annotations are obtained by user study, not all of the samples have ranked 3CADs. Therefore, we filter the annotation samples whose number of ranked CAD models is less than 3 CADs to be able to see the effect of clean and complete dataset. In that way, we increase Top-1 Retrieval Accuracy by %3 to %6.

Method	Margin (neg,pos)	trash bin	bathtub	bed	bookshelf	cabinet	chair	display	file	sofa	table	other	class avg	inst avg
Baseline	$m_{triplet} = 0.2$	0.50	**0.60**	0.42	0.19	0.26	0.55	0.45	0.25	0.33	0.32	0.43	0.39	0.43
Random neg sampling	$m_n = 1.0$	0.68	0.40	0.45	0.59	0.29	0.50	0.40	0.50	0.48	0.49	0.44	0.47	0.50
Random neg sampling	$m_n = 1.25$	**0.86**	0.33	**0.61**	0.56	0.32	0.51	0.42	0.42	0.50	0.42	**0.56**	0.50	0.52
Random neg sampling	$m_n = 1.50$	0.63	0.53	0.30	**0.71**	0.42	0.47	0.42	0.50	0.43	0.48	0.32	0.47	0.49
Random neg sampling	$m_n = 1.25, m_p = 1.25$	0.52	0.47	0.45	0.37	0.42	0.50	0.50	0.42	0.46	0.42	0.38	0.45	0.46
Random neg sampling	$m_n = 1.25, m_p = 0.1$	0.72	0.27	0.54	0.28	**0.51**	0.44	0.50	0.25	0.46	0.48	0.54	0.45	0.49
Same and diff. class	$m_n = 1.25, 0.1, m_p = 0$	0.55	0.53	0.27	0.47	0.39	0.54	0.42	0.50	0.46	0.52	0.47	0.46	0.50
Same and diff. class	$m_n = 1.25, 0.2, m_p = 0$	0.47	0.53	0.58	0.49	0.35	**0.56**	0.43	0.25	0.41	0.54	0.46	0.46	0.50
Same and diff. class	$m_n = 1.25, 0.3, m_p = 0$	0.45	0.53	**0.61**	0.42	0.41	0.55	0.42	0.33	0.52	**0.58**	0.44	0.48	0.50
Random sampling&reg	$m_n = 1.0 m_p = 0$	0.79	0.40	0.45	0.46	0.50	0.49	0.48	**0.58**	**0.52**	0.47	0.44	0.51	0.51
Random sampling&reg	$m_n = 1.25 m_p = 0$	0.81	0.53	0.58	0.61	0.45	0.50	**0.52**	0.42	0.46	0.44	0.43	**0.52**	**0.53**

Table 5. Top1 Retrieval Accuracy comparison of baseline paper [6] and our contrastive loss setting per class on test split of Scan-CAD Object Similarity benchmark public dataset.