Single Image 3D Shape Retrieval via Cross-Modal Instance and Category Contrastive Learning – *Supplementary Material*–

Ming-Xian Lin[1,3] **Jie Yang**[1,3] **He Wang**[4] **Yu-Kun Lai**[5]
Rongfei Jia[6] **Binqiang Zhao**[6] **Lin Gao**[1,2,3*]
[1] Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences
[2] Zhejiang Lab [3] University of Chinese Academy of Sciences
[4] Center on Frontiers of Computing Studies, Peking University
[5] School of Computer Science and Informatics, Cardiff University
[6] Tao Technology Department, Alibaba Group
{linmingxian20g, yangjie01, gaolin}@ict.ac.cn hewang@pku.edu.cn
LaiY4@cardiff.ac.uk {rongfei.jrf, binqiang.zhao}@alibaba-inc.com

Abstract

In this supplemental material, we provide comparison with the SOTA generic cross-modal retrieval method, more ablation studies, more visualization, and a discussion on future work.

1. Result Comparison with the Generic Cross-modal Retrieval Method

Jing *et. al.* [2] were recently published in CVPR 2021, which focused on generic cross-modal retrieval and proposed a new loss, *i.e.* cross-modal center loss, to learn discriminative and modal-invariant features for data from different modalities. During training, they propose to map different modalities *e.g.* mesh, multi-view images, and point clouds of the instance into features with small intra-class variation across all modalities. Although their task is not the same as ours, the work provides results on single image shape retrieval from a synthetic dataset, ModelNet40 [4]. We therefore compare our method with this state-of-the-art generic cross-model retrieval work on the Pix3D dataset. The results can be seen in Table 1.

Experiment setting. We train [2] in Pix3D only selecting the constraints on meshes and images in [2] without considering the constraints on point clouds. For a fair comparison, the image size is 224×224; the batch size is 60; the learning rate is 5e-5; the methods of data augmentation are the same as our proposed approach,

Dataset	Method	Acc_{Top-1}
Pix3D	ours proposed approach	**78.9%**
Pix3D	Cross-modal Center Loss [2]	62.4%

Table 1. **Results of Top-1 retrieval accuracy on Pix3D [3] compared to Cross-modal Center Loss [2].** Cross-modal center loss aims to learn discriminative and modal-invariant features for data from different modalities. It is shown that our proposed method has a better performance on the Pix3D dataset, which consists of real images.

Result. We can see that our proposed approach is about 16% better than [2] in Pix3D on the accuracy of Top-1 retrieval. We think the reason for performance degradation of [2] in Pix3D is that the query images in the Pix3D dataset are real images while the images in ModelNet40 [4] used by [2] are synthetic. The gap between different query images of the same instance is relatively large in Pix3D, which makes the center loss more difficult to converge. Our proposed method does not require the features of the query image and rendered images to be close to one central feature. Instead, we generate a query-specific feature for the rendered image despite the fact that different modalities have different features.

2. More Ablation Studies

The setting of β_1. β_1 controls the relative weight when combining the instance and category losses. The accuracy at the instance level will be affected if the proportion of cat-

*Corresponding Author is Lin Gao

egory loss is too large; however, the category loss will not have sufficient effect if its proportion is too small. Therefore, the setting of β_1 in Eq. 6 has an impact on the experiment results. We gradually increase β_1 from 0.0 to 1.0 on the Pix3D dataset. The results shown in Fig 1 confirm our points above. The category loss can help the network improve the retrieval accuracy at the instance level through category information when β_1 is less than 0.2. However, the category loss has a negative impact on the accuracy of retrieval at the instance level by improving the similarity of other same category samples of the query image too much when β_1 is greater than 0.2. Therefore, we choose 0.2 as the value of β_1.

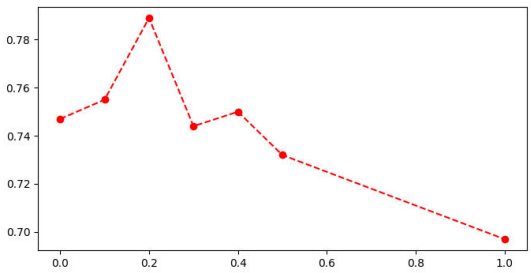

Figure 1. **Impact of β_1.** The x-axis indicates the setting of β_1 and the y-axis indicates the accuracy of top-1 retrieval on Pix3D.

The usage of masks. The mask can make the encoder focus more on the objects that need to be retrieved in the query image. There are three typical ways regarding a mask, including concatenating the query image with the mask image, multiplying the query image with the mask image element-wise and using the query image without the mask image. Therefore, we have explored the usage of the mask here to find a suitable way. The results shown in Table 2 indicate that concatenating the query image with the mask image is the best way for the IBSR task in our proposed pipeline. The first method can retain more information to prevent the pre-generated mask from being of insufficient quality compared to the second method and can focus more on the object to be retrieved compared to the third method.

The number of views and the rendering method. we perform more ablation studies for the parameters in Eqs.1-6, as shown in Table 3, including the number of views and the rendering method. For the number of views, when it is smaller than 12, increasing it can be helpful. When it is greater than 12, increasing it brings less benefit but more training burden. For the rendering method, the results with the basic rendering pipeline are not satisfactory. We believe that this introduces a prior about the texture for 3D shapes,

Dataset	Usage	Acc_{Top-1}
Pix3D	as another channel	**78.9%**
Pix3D	masked	29.6%
Pix3D	without mask	62.1%

Table 2. **Results of ablation studies on the usage of the mask of query image**. 'as another channel' means the input of query image encoder is the concatenation of the query image and the mask image; 'mask' means the input of query image encoder is the query image multiplied with the mask image element-wise; 'without mask' means the input of query image encoder is the query image only. It is shown that 'as another channel' can provide more information compared to 'mask' and focus on the object to be retrieved compared to 'without mask'.

Params	$\text{Acc}_{\text{Top}_1}$				
the number of views	1	6	12	18	24
	55.6%	**68.3%**	**78.9%**	**79.3%**	**79.2%**
rendering method	grayscale		basic rendering pipeline		
	78.9%		**62.8%**		

Table 3. **Additional ablation study on Pix3D.**

which causes the network to be distracted from learning the geometry features.

The attention module & the scalability. We got 62.1% for $\text{Acc}_{\text{Top}_1}$ on Pix3D when removing the attention module. Therefore, it is unavoidable to iterate over the full set of 3D shapes for each query in our method. However, the time spent by our method is acceptable in most of the current 3D datasets. When loaded onto an RTX 3090 GPU in advance, the retrieval on Pix3D (322 shapes) takes about 15ms and on ShapeNet chair (6778 shapes) takes about 17ms for each query when grayscale images of 3D shapes are used.

3. Visual Comparison of Color Augmentation Methods

We introduce the color transfer mechanism into contrastive learning, which is a more powerful color augmentation that improves the robustness of the network, helping the network extract color-independent features. We also compare the visual results of color transfer and random HSV augmentation, as shown in Fig. 2. From the results, we can see that there are two main advantages for color transfer compared to HSV augmentation. On the one hand, the color transfer can augment query images using another query image as a reference, which decouples the object and color in 2D images. On the other hand, the results of the color transfer mechanism look more natural than HSV augmentation because the color transfer can extract color statistics from real pictures while HSV augmentation is a random perturbation.

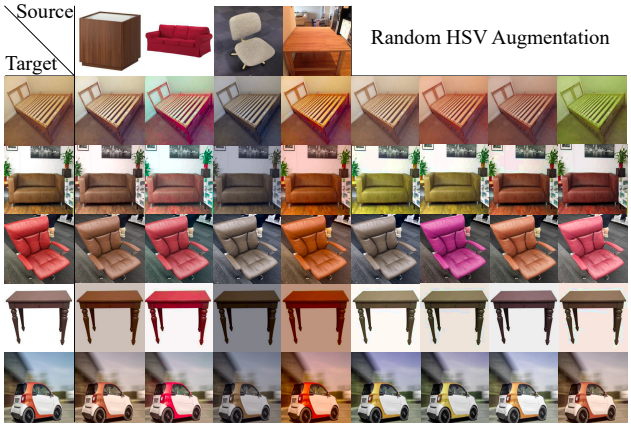

Figure 2. **The results of color transfer and random HSV augmentation.** 'Source' means the source color image and 'Target' means the target shape image. The columns from 1 to 4 are the results of color transfer, and the columns from 5 to 8 are the results of random HSV augmentation. Random HSV augmentation means that we random shift hue, saturation and value of query images.

4. Future Work

In the ablation study about the mask, we showed the effect of the mask on accuracy. Our method currently is to apply the mask by Mask R-CNN [1] and OCRNet [5]. In future work, we will integrate the mask generation into our proposed framework.

References

[1] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 4323

[2] Longlong Jing, Elahe Vahdani, Jiaxing Tan, and Yingli Tian. Cross-modal center loss for 3D cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3142–3151, June 2021. 4321

[3] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B. Tenenbaum, and William T. Freeman. Pix3D: Dataset and methods for single-image 3D shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 4321

[4] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 4321

[5] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 173–190, Cham, 2020. Springer International Publishing. 4323