Towards Cross-Modal Point Cloud Retrieval for Indoor Scenes

Fuyang Yu[1], Zhen Wang[2], Dongyuan Li[2], Peide Zhu[3], Xiaohui Liang[1(✉)], Xiaochuan Wang[4], and Manabu Okumura[2]

[1] Beihang University, Beijing, China
liang_xiaohui@buaa.edu.cn
[2] Tokyo Institute of Technology, Tokyo, Japan
[3] Delft University of Technology, Delft, Netherlands
[4] Beijing Technology and Business University, Beijing, China

Abstract. Cross-modal retrieval, as an important emerging foundational information retrieval task, benefits from recent advances in multimodal technologies. However, current cross-modal retrieval methods mainly focus on the interaction between textual information and 2D images, lacking research on 3D data, especially point clouds at scene level, despite the increasing role point clouds play in daily life. Therefore, in this paper, we proposed a cross-modal point cloud retrieval benchmark that focuses on using text or images to retrieve point clouds of indoor scenes. Given the high cost of obtaining point cloud compared to text and images, we first designed a pipeline to automatically generate a large number of indoor scenes and their corresponding scene graphs. Based on this pipeline, we collected a balanced dataset called CRISP, which contains 10K point cloud scenes along with their corresponding scene images and descriptions. We then used state-of-the-art models to design baseline methods on CRISP. Our experiments demonstrated that point cloud retrieval accuracy is much lower than cross-modal retrieval of 2D images, especially for textual queries. Furthermore, we proposed ModalBlender, a tri-modal framework which can greatly improve the Text-PointCloud retrieval performance. Through extensive experiments, CRISP proved to be a valuable dataset and worth researching. (Dataset can be downloaded at https://github.com/CRISPdataset/CRISP.)

Keywords: Point Cloud · Cross-modal Retrieval · Indoor Scene

1 Introduction

With the advancement of multimodal technologies, various tasks are now benefiting from multimodality. Unlike traditional information retrieval methods, cross-modal retrieval [26] involves not only text, but also other modalities, such as images and videos, in both queries and retrieved results. Currently, the most

F. Yu and Z. Wang—Equal contribution.

S. Rudinac et al. (Eds.): MMM 2024, LNCS 14557, pp. 89–102, 2024.
https://doi.org/10.1007/978-3-031-53302-0_7

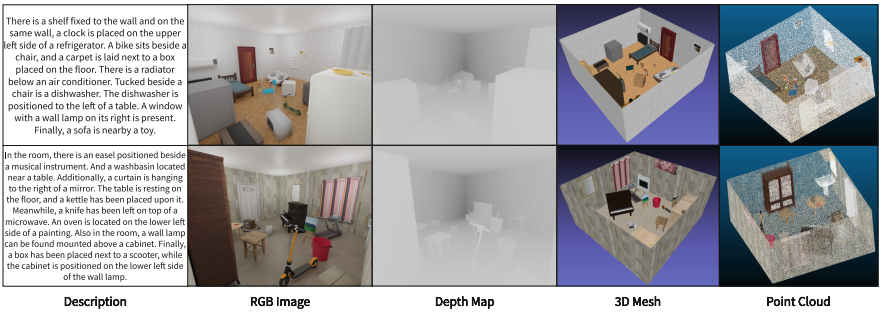


Fig. 1. Two samples in CRISP.

studied cross-modal retrieval is the retrieval between text and images [13,16]. With the development of cross-modal methods, BLIP [9] has even achieved close to 100% @1 accuracy on the Flickr30k [16] dataset. Compared to 2D data, 3D data at scene level can provide more spatial information and is less susceptible to occlusion. Despite the development of 3D technology driven by deep learning, there is still a lack of exploration of 3D data in cross modal tasks. 3D data retrieval can establish a bridge between 3D data and other modal data, will be a critical tool of the effective management and utilization of massive 3D data in the future. However, this fundamental task is constrained by the lack of datasets and has developed slowly. Although some previous work [4,20,22] has explored real indoor scenes with RGB-D or point cloud, the existing indoor scene point cloud datasets still have the following issues. a) Small size, the most widely used Scannet [4] dataset only contains around 1,500 scans of different rooms, whereas the text-image cross-modal retrieval dataset Flickr30k has a test set with 1k samples. b) Suffering from long tail distribution, the unbalanced distribution of object categories in dataset makes the models easy to learn the bias.

To address these issues, inspired by some 2D image synthetic datasets such as CLEVR [8] and CLEVRER [29], we designed a novel pipeline to automatically compose a **balanced** large number of realistic 3D indoor scenes with scanned objects and collect point clouds based on them. We use synthetic technology to build our dataset because it is low cost and can be easily scaled to larger numbers while being more balanced without bias [8]. To collect textual description and 2D photographs for the query, we recorded the generated scene information and created a scene graph for each point cloud, with which we constructed the *textual description* of the scene. We also took a photograph from a random corner of a room to obtain a *RGB image* along with its *depth map*. With these three modalities, we finally constructed the **CRISP** (**C**ross-modal **R**etrieval on **I**ndoor **S**cenes **P**oint-cloud) dataset. Two examples of CRISP are shown in Fig. 1.

To our knowledge, CRISP stands as the first and largest balanced indoor dataset that facilitates point cloud data retrieval, either through scene descriptions or rendered images. This dataset serves as a robust benchmark, offering researchers a valuable tool to assess the efficacy of their models. To demonstrate

the usefulness and reasonability of our dataset, we leveraged some state-of-the-art methods to establish strong baselines for CRISP, including Text-PointCloud retrieval and Image-PointCloud retrieval. Specifically, we proposed a new framework called ModalBlender, which brings cross-modal attention and intermediate-modal alignment into Text-PointCloud retrieval and greatly improves the overall performance while maintaining retrieval efficiency, providing a novel and useful approach to enhance text retrieval of point clouds. Our experiments showed that CRISP is challenging and, together with our baselines, provides a strong starting point for future research in indoor scene understanding.

Our main contributions in this study can be summarized as follows: (1) We proposed a general and flexible pipeline that utilizes existing 3D object models to generate a large quantity of photorealistic indoor scenes together with the scene graph; (2) Based on this pipeline, we constructed a dataset called CRISP, which contains massive of text, images, and point cloud that makes it now possible to research cross-modal point cloud retrieval; (3) We evaluated existing SOTA methods and proposed a new framework ModalBlender to further improve the retrieval performance and proved its validity through detailed experiments.

2 Related Work

2.1 Indoor Scene Datasets

➢ **Real Scanned Datasets:** One of the earliest indoor scene datasets is NYU Depth Dataset V2 [20], which contains RGB-D data of indoor scenes with labeled objects and semantic segmentation masks. The SUN RGB-D [22] dataset is another popular dataset, which includes RGB-D data of indoor scenes with semantic annotations, object instances, and scene categories. The Scannet [4] dataset is a large-scale indoor scene dataset that includes both RGB-D and point cloud data of indoor scenes. The Matterport3D [1] dataset provides high-quality RGB-D data and 3D panoramic of large-scale indoor scenes with object instances and semantic annotations.

➢ **Synthetic Scene Datasets:** Compared to scanning scenes, synthetic datasets offer the advantage of easy scalability to larger amounts of data. SceneNet [6] comprises a variety of annotated indoor scenes that have been widely utilized for object detection, semantic segmentation, and depth estimation. SUNCG [23] contains numerous diverse indoor scenes with highly detailed object annotations, making it useful for 3D reconstruction and semantic parsing. InteriorNet [11] is a dataset featuring photo-realistic indoor scenes, which is useful for evaluating methods on real-world data.

2.2 3D Retrieval Datasets

Previous 3D retrieval related dataset and research is mostly about retrieving single 3D objects using 2D images. Some of the popular datasets include the Princeton Shape Benchmark (PSB) [19], the ShapeNet [2], ModelNet40 [27], and MI3DOR [21]. The PSB contains a large collection of 3D models with varying

Table 1. Comparison of CRISP with other indoor scene datasets. Our dataset is currently the only one that suitable for benchmark cross modal point cloud retrieval.

Dataset	Sample	Room	Obj Cate.	Type	RGBD	Desc.	Scene Graph	Point Cloud
Scannet [4]	1,513	707	21	Real	✓	✗	✗	✓
NYU-Depth [20]	1,449	464	13	Real	✓	✗	✗	✗
SUN-RGBD [22]	10,335	10,335	800	Real	✓	✗	✗	✗
3DSSG [25]	1,482	478	160	Real	✓	✗	✓	✓
SUNCG [23]	45,622	45,622	84	Syn	✓	✗	✗	✗
InteriorNet [11]	5M	1.7M	158	Syn	✓	✗	✗	✗
SceneNet [6]	10,030	57	13	Syn	✓	✗	✗	✗
CRISP (Ours)	10,000	10,000	62	Syn	✓	✓	✓	✓

levels of complexity, while ShapeNet and ModelNet40 aim to provide large-scale objects with different annotated categories. The MI3DOR dataset offers monocular image-based 3D object retrieval. These datasets have been used extensively in the literature to test and compare different 3D object retrieval algorithms and have led to significant advancements in the field. The only related 3D *scene* retrieval dataset is SHREC'19 [30], which uses one 2D sketch-based image to retrieve 3D scenes with 30 categories such as library and supermarket.

3 CRISP Dataset

As shown in Table 1, existing indoor scene datasets are divided into real and synthetic types. Some commonly used real-scene datasets like Scannet [4] have a small number of room scans, limited by the difficulty of data collection, while synthetic datasets can have larger numbers of scenes. Additionally, most of these datasets do not consider data distribution and often exhibit long-tail effects. The scarcity of examples for rare categories hinders models from learning robust 3D features. To our knowledge, CRISP is the **earliest** and **largest balanced** indoor retrieval dataset that facilitates exploration of the interaction between 3D data with text or image data. Figure 1 presents examples from the dataset. The creation of CRISP began with the collection of objects and hierarchical scene generation. Throughout the generation phase, our utmost priority was to maintain dataset equilibrium, thereby minimizing potential learning biases within models trained on the dataset. Using the generated scenes, we collected distinct sets of point cloud data, image data, and textual description data, all amalgamated to form the comprehensive CRISP dataset as shown in Fig. 4.

3.1 Object Collection

We first selected 62 common seen indoor object categories into our dataset which can encompass the majority of indoor objects encountered in daily life and meanwhile guarantee the scene diversity. Each object category contains 1∼5 different

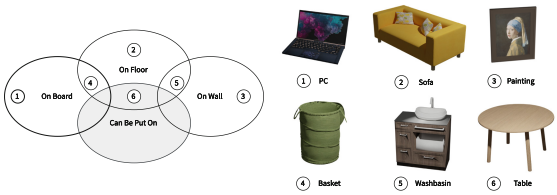


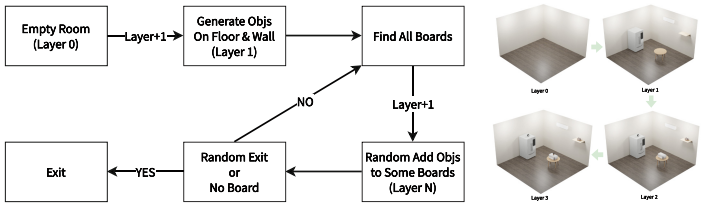
Fig. 2. This is the Venn diagram we used to classify object properties. The light-colored ovals represent the positional attributes of the objects, namely board, floor, and wall. Board refers to any flat surface other than the floor. The gray ovals indicate whether an object can have something placed on top of it. (Color figure online)

instances. Totally we collected 171 different instances. We aimed to create a well-balanced dataset while also ensuring that the scenes composed of these objects are more realistic. To achieve the second goal, we analyzed the layout of indoor scenes and categorized the objects accordingly. Inspired by some automated facilities layout technologies [12], we determined the common locations of each type of object in the scene and sorted them into the following categories: "Object On Board", "Objects On Floor", "Objects on Wall", as illustrated in Fig. 2. Those categories were established based on the spatial positional relationships of objects. Following the classification of an object's spatial position, we further categorized it according to the presence of a support surface that can hold other objects, as attribute "Can Be Put On". By combining these attributes, we can ensure that the generated scenes are reasonably plausible.

3.2 Hierarchical Generation

We used Kubric [5] library to load and organize objects. During the generation process, we employed a hierarchical generation method that incorporates the object category information obtained in Sect. 3.1 to achieve a higher degree of realism, i.e., one indoor scene should contain various categories of objects, as well as rational horizontal and vertical spatial object relationships. Detailed steps are shown in Fig. 3(a) and an example is shown in Fig. 3(b).

Hierarchical generation can combine single simple objects into a complex indoor scene and ensure rich spatial relationship between objects, clearly display the dependency relationships between objects, which facilitates the generation of scene graph. We maintained a global 3D occupancy map that records the space occupied by objects to prevent collisions during generation. Meanwhile, we maintained the randomness of object categories and positions, making the dataset more balanced and free from long-tailed distributions, providing a more fair and robust benchmark for further research. The generated scene was collected in the format of **3D mesh** for subsequent generation.



(a) Hierarchical generation procedure. (b) Example.

Fig. 3. Our hierarchical generation procedure begined with layer 0, which represents an empty room. We then proceed to layer 1, where we added objects to the floor and walls. Next, we identified all available boards, and randomly place some objects on them, which became the next layer. This step was repeated until a random exit was triggered or there are no more boards available.

3.3 Point Cloud Collection

We converted the 3D scene mesh generated in Sect. 3.2 to ".obj" format and used CloudCompare[1], a 3D point cloud processing software, to sample point cloud. We sampled 200,000 points on surface which ensuring the collected point cloud much closer to the actual collected, and adjusted the rendering effect to make the collected point cloud as similar as possible to the object model.

3.4 Scene Image Collection

We used the bpy library provided by Blender[2] to generate photorealistic images from pre-built indoor scenes. For each room, we randomly selected a corner and simulated the breadth and height of human vision using a camera to capture the corresponding RGB image. We took one photograph for each room. To achieve more realism, we adjusted the rendering effect of Blender and added point light sources of different energy to simulate changes in lighting that occur in real scenes. We also extracted the depth image corresponding to the RGB image collected above to provide more information and support further research.

3.5 Scene Graph Generation

With the help of controllable generation procedure, we were able to record lots of useful information during scene generation such as the precise coordinate position of each object and where they were put on. We use this information to generate the scene graph.

In our designed scene graph, the nodes represent different objects, while the edges represent the spatial relationships between the nodes. Several types of relationships are defined, including "Next To", "Support" and "Positional Realation".

[1] http://www.cloudcompare.org/.
[2] https://www.blender.org/.

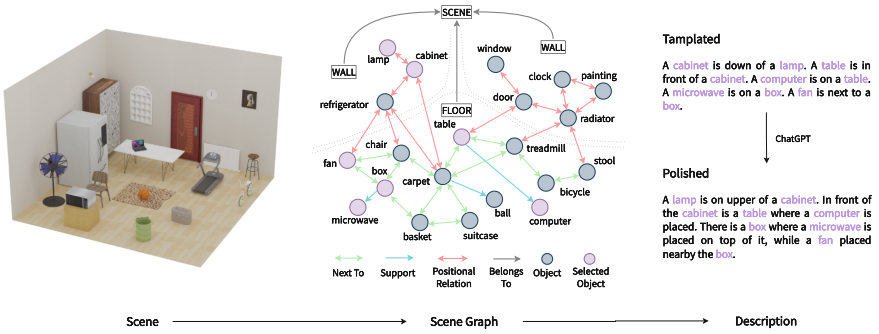


Fig. 4. Pipeline of the generation of CRISP, including scene generation (Sect. 3.2), scene graph generation (Sect. 3.5) and discription generation (Sect. 3.6). (Color figure online)

When two object are placed on the same surface and close enough, they form a "Next To" relationship. When one object is placed on the other object, these two object form a "Support" relationship. When for two objects, there are at least one object is on a wall, then they compose a "Positional Relation". The generation of scene graph transformed the layout of the scene into a form that the computer can directly process. An example can be found in Fig. 4. Compared to human annotation, our scene graph generation is fast while ensuring the correctness.

3.6 Textural Description Collection

After scene graph generation, as shown in Fig. 4, a random subset of nodes and edges was selected to form a description. The template to generate the descriptions is shown in the following prompt "<Object_A> is <R> <Object_B>", where "<Object_A>" and "<Object_B>" represent the categories of different nodes, and "<R>" denotes a spatial relationship between the two nodes. After generating the templated descriptions, we utilized the OpenAI ChatGPT API[3] to optimize them and make them more consistent with human language conventions. Once the rewriting was finished, we employed manual check by human to proofread the description meticulously, ensuring that it will adhere to the original meanings without mistakes and obtained the final textual descriptions.

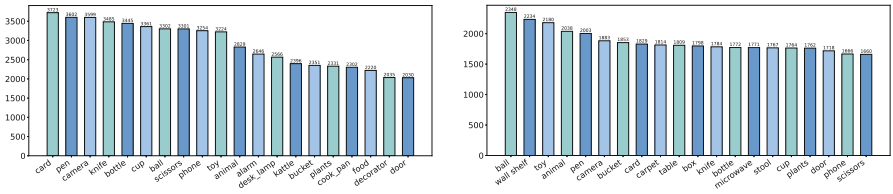
3.7 Dataset Statistics

CRISP comprises 10,000 scenes, each including one or three descriptions, a set of paired RGB and depth images, and a point cloud of the scene. The detailed statistic is shown in Table 2. On average, a scene in CRISP contains 27.36 objects. We then analyzed the frequency of object category occurrences and Fig. 5 illustrates the top 20 most commonly found categories in point cloud or description.

[3] https://platform.openai.com/docs/models/gpt-3-5.

Table 2. Numerical Statistics of CRISP.

Modal	# Train	# Val	# Test	# Total
Text	24,000	1,000	1,000	26,000
Image	8,000	1,000	1,000	10,000
Point Cloud	8,000	1,000	1,000	10,000



(a) Average objects in scenes. (b) Average objects in descriptions.

Fig. 5. Distribution of top-20 object categories in CRISP.

It can be seen that both the occurrence frequency of objects in the generated scenes and in the descriptions is almost the same, indicating a well-balanced dataset compared to other indoor datasets like Scannet [4]. The characteristic of CRISP enables models trained on it to focus more on extracting different modal features rather than learning biases.

3.8 Unique Features of CRISP

We summarize the highlights of CRISP as follows: **(1) Easy to expand.** We developed an automatic and highly efficient pipeline for generating synthetic indoor scenes and collecting data in various modals, which allows us to easily expand the size of CRISP, and can be extended to construct datasets for other tasks. **(2) Large-Scale.** CRISP is currently the largest point cloud dataset for cross-modal exploration with 10,000 scenes. The extensive data in CRISP ensure a more diverse and representative set of multi-modal data, enabling more accurate and robust models. **(3) Without small data bias.** As shown in Sect. 3.7, our dataset exhibits a strong balance in terms of object category distribution, scene variability, and textual description diversity , ensuring that there is no small data bias. **(4) Novelty and broader applicability.** CRISP is a novel and versatile cross-modal retrieval dataset, making it a cutting-edge and highly promising development in cross-modal retrieval. Its multimodal nature enables various applications such as scene understanding, and more, allowing for future research not only in retrieval, but also other multimodal point cloud tasks.

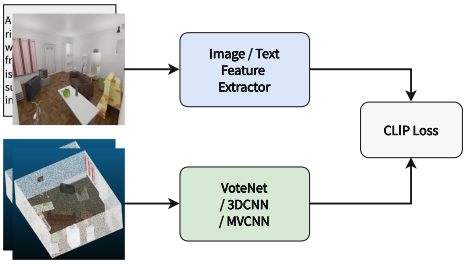


Fig. 6. The Pipeline of we used two-stream cross-modal retrieval baseline model.

4 Baseline Methods

4.1 Model Architecture

Considering the exceptional performance of models based on the CLIP architecture [9,10,18] on the MSCOCO [13] and Flickr [16] datasets, to create a strong baseline, in this paper, we adopted a similar dual-stream model with CLIP loss, see Formula (1). The pipeline is shown in Fig. 6. One stream of the model consists of queries, which in this case are either images or textual descriptions of rooms, while the other stream comprises queried values, specifically point cloud of the rooms. State-of-the-art methods were utilized to extract features from the inputs of both streams, and the CLIP loss, was then applied to align the features of the two different modalities by calculating the loss.

$$L_C = \frac{1}{2}\left[\frac{1}{B}\sum_{k=1}^{B}\frac{\exp\left(S_{v_k,t_k}/\tau\right)}{\sum_l^B \exp\left(S_{v_k,t_l}/\tau\right)} + \frac{1}{B}\sum_{k=1}^{B}\frac{\exp\left(S_{v_k,t_k}/\tau\right)}{\sum_l^B \exp\left(S_{v_l,t_k}/\tau\right)}\right], \tag{1}$$

Where B is the batch size and τ is the temperature hyper-parameter. $S_{v,t}$ is the total similarity score, defined as Formula (2):

$$S_{v,t} = \frac{1}{2}\left(\sum_{i=1}^{N_v} w_v^i \max_j a_{ij} + \sum_{j=1}^{N_t} w_t^i \max_i a_{ij}\right). \tag{2}$$

The variables u and v represents two different modalities, such as text and point clouds, while a_{ij} is a feature similarity matrix obtained by multiplying features from different modalities. The weights of the modal features are represented by w_v^i and w_t^j, obtained by $\left[w_v^0, w_v^1, ..., w_v^{N_v}\right] = \text{Softmax}\left(\text{MLP}_v\left(V_f\right)\right)$, MLP_v represents the fully connected layers used to encode modal v.

After training and before testing, for each kind of data in test set, we used its corresponding feature extractor to precalculate the feature of each sample, and then when testing, we would directly use these features to calculate the cosine similarity for retrieving to ensure the retrieval efficiency.

Table 3. Experimental Result on CRISP. The parentheses under the "InferTime" indicate the time taken for preprocessing MVCNN, which involves generating multi-view images for 1000 point cloud scenes. "w/o" means "without". Recall is used as an evaluation matrix. One out of a thousand candidates is chosen as the result, and its correctness is noted as $R@1$, and $R@5$, $R@10$ and $R@100$ are defined in the same way.

Query	**PC Model**	**R @1**	**R @5**	**R @10**	**R @100**	**# Param**		**InferTime**
						Train	**Test**	
Rand	-	0.1	0.5	1.0	10.0	–	–	–
Cross-Modal Retrieval								
Image	VoteNet	43.8	90.5	97.4	99.8	89M	89M	77 s
Image	3DCNN	17.4	46.8	62.5	99.1	136M	136M	179 s
Image	MVCNN	92.5	99.4	99.7	100	188M	188M	144 s (+86m)
Text	VoteNet	0.1	0.5	1.0	10.0	134M	134M	73 s
Text	3DCNN	1.5	4.7	7.9	37.7	172M	172M	159 s
Text	MVCNN	0.5	2.9	5.6	42.7	213M	213M	144 s (+86m)
ModalBlender								
Text	VoteNet	1.1	3.0	6.2	37.8	222M	134M	73 s
Text	3DCNN	5.6	13.3	20.7	55.1	261M	172M	159 s
Text	MVCNN	8.7	28.5	40.9	89.5	302M	213M	144 s (+86m)
w/o CMM	MVCNN	6.9	20.3	31.4	72.3	–	–	–
w/o IMA	MVCNN	2.1	5.7	10.2	58.7	–	–	–

4.2 Implementation Details

We utilized pretrained Swin Transformer [15] and RoBERTa [14] to extract features from images and scene descriptions separately. To extract features from point clouds and compare pros and cons of different 3D approaches, we employed three different methods. The first method, VoteNet [17], which extracted features directly from the point cloud data. The second method, 3DCNN [28], first voxelized the point cloud and then used sparse convolution to extract features, here we used our own designed sparse convolution network with Spconv [3] which had a model architecture similar to ResNet-50 [7]. The third method, MVCNN [24], captured point cloud information from different angles by taking snapshots, which were then used to extract features. For MVCNN, we also used pretrained Swin Transformer as the backbone.

4.3 Experimental Result

The experimental results of the cross-modal retrieval are shown in Table 3.

For **Image-PointCloud retrieval**, MVCNN performed the best, far surpassing VoteNet and 3DCNN. This was because MVCNN converts 3D point cloud features to 2D pixel features by rendering the point cloud. Retrieving 2D

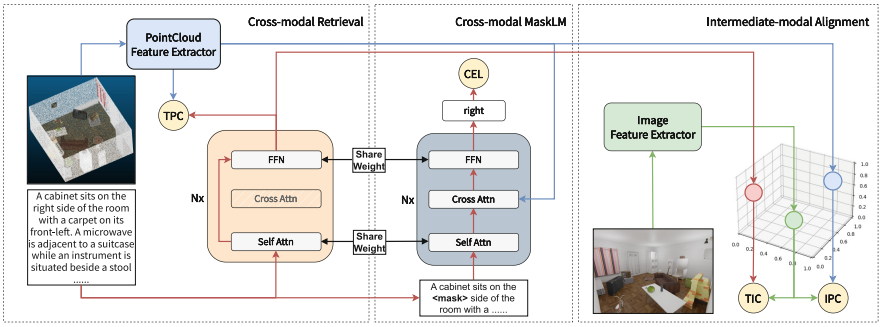


Fig. 7. Pipeline of our proposed ModalBlender.

images using 2D image search benefits from high-performance pretrained models. However, compared to VoteNet and 3DCNN, which can process point cloud data directly, MVCNN must first render and photograph point cloud data, making it much slower. It took around 86 min to preprocess all the CRISP test set for MVCNN before testing. Additionally, because MVCNN typically used a fixed camera angle to photograph the rendered result, its performance was greatly affected when there were obstructions, such as a ceiling in a room. On the contrary, VoteNet and 3DCNN were more robust. How to improve model accuracy and maintaining efficiency remains a worthy research question.

While for **Text-PointCloud retrieval**, all three models performed poorly. VoteNet performed the worst and resulted in an accuracy similar to random selection. 3DCNN, which was based on 3D data and 3D convolution kernels, performed the best in Text-PointCloud retrieval, because it had better spatial perception than MVCNN and VoteNet. Overall, the accuracy of all three backbones was very low because of the huge feature gap between text and point cloud. Meanwhile, currently there was no widely-used pretrained model to help Text-PointCloud alignment, making the situation worse. CRISP as the first cross-modal retrieval dataset for text and point cloud, provides a convenient and useful benchmark for studying the alignment of text and point cloud modalities.

5 Text Point-Cloud Alignment

5.1 Architecture and Implementation

To address the challenge of aligning textual and point cloud features, we then proposed a model called **ModalBlender** which comprises three submodules, as shown in Fig. 7. The first submodule, **CMR** (Cross-modal Retrieval), was the same as the one described in Sect. 4.1, and its output loss is Text-PointCloud CLIP loss (TPC). The second submodule, **CMM** (Cross-modal MaskLM), was a model similar to RoBERTa, but with cross-attention layers that take text feature (Q) and point cloud feature (K and V) as inputs. The self-attention layers in CMR and CMM share the same weights. We randomly masked 15% of the original text query and use it as input to CMM, and the output loss

of CMM is cross-entropy loss (CEL). We adopt cross attention mechanism on text and point cloud features to guide the understanding of each other. The third submodule, **IMA** (Intermediate-modal Alignment), used two additional CLIP losses that involve image features as the intermediate modal to align text and point cloud. We paired text and image features to calculate Text-Image CLIP loss (TIC) and paired image and point cloud features to calculate Image-PointCloud CLIP loss (IPC). We leveraged image feature as the intermediate modal because there are well-developed text and image pretrained models and also image has RGB information that can be directly mapped into point cloud data, which made image a perfect bridge modal to align text and point cloud features. Finally, we used the weighted sum of these four losses as the final loss of ModalBlender. In particular, when testing, only the CMR module was used.

5.2 Quantitative Analysis

➢ **Performance Comparison:** In Table 3, we present the experimental results of ModalBlender. The performance of all three different backbone models had been significantly improved, indicating the effectiveness of the CMM and IMA modules. MVCNN achieved the highest accuracy and the greatest improvement, followed by 3DCNN. Although VoteNet still had the lowest accuracy among the three models, it was able to show some effective accuracy. The experimental results strongly demonstrated that cross-modal attention and the use of image features as an intermediate modal could greatly facilitate alignment between text and point cloud modalities.
➢ **Ablation Study:** We then conducted ablation experiments on ModalBlender with MVCNN. The results in the last two rows of Table 3 show that the removal of the CMM or IMA module degraded the performance. This demonstrated that both CMM and IMA are important in improving the model's accuracy. And compared to CMM, IMA has a greater influence on overall performance.

6 Conclusion

In this paper, we introduced CRISP, the first 3D indoor balanced scene cross-modal retrieval dataset that focuses on retrieving point cloud using text or images. Given the difficulty of obtaining 3D point cloud data, we proposed an automated pipeline that can generate a vast number of realistic indoor 3D point cloud scenes and then formed a dataset called CRISP cantains point clouds, RGBD-images and textual descriptions of the scenes. We conducted comprehensive experiments based on CRISP using now SOTA methods, and observed a huge performance gap between CRISP tasks and previous Text-Image retrieval tasks especially when using text. As one step towards better Text-PointCloud retrieval, we proposed a novel architecture named ModalBlender, and experimental evidence demonstrated that ModalBlender could significantly improve the accuracy of retrieval and provide a useful approach for aligning text and point cloud features in the absence of pretraining.

References

1. Chang, A., et al.: Matterport3D: learning from RGB-D data in indoor environments. arXiv preprint arXiv:1709.06158 (2017)
2. Chang, A.X., et al.: ShapeNet: an information-rich 3D model repository. arXiv preprint arXiv:1512.03012 (2015)
3. Contributors, S.: SpConv: spatially sparse convolution library. https://github.com/traveller59/spconv (2022)
4. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: richly-annotated 3D reconstructions of indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5828–5839 (2017)
5. Greff, K., et al.: Kubric: a scalable dataset generator. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3749–3761 (2022)
6. Handa, A., Pătrăucean, V., Stent, S., Cipolla, R.: SceneNet: an annotated model generator for indoor scene understanding. In: 2016 IEEE International Conference on Robotics and Automation (ICRA), pp. 5737–5743. IEEE (2016)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
8. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2901–2910 (2017)
9. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
10. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning, pp. 12888–12900. PMLR (2022)
11. Li, W., et al.: InteriorNet: mega-scale multi-sensor photo-realistic indoor scenes dataset. arXiv preprint arXiv:1809.00716 (2018)
12. Liggett, R.S.: Automated facilities layout: past, present and future. Autom. Constr. **9**(2), 197–215 (2000)
13. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
14. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
15. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
16. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2641–2649 (2015)
17. Qi, C.R., Litany, O., He, K., Guibas, L.J.: Deep hough voting for 3D object detection in point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9277–9286 (2019)

18. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
19. Shilane, P., Min, P., Kazhdan, M., Funkhouser, T.: The Princeton shape benchmark. In: Proceedings Shape Modeling Applications, 2004, pp. 167–178. IEEE (2004)
20. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. ECCV **5**(7576), 746–760 (2012). https://doi.org/10.1007/978-3-642-33715-4_54
21. Song, D., Nie, W.Z., Li, W.H., Kankanhalli, M., Liu, A.A.: Monocular image-based 3-D model retrieval: a benchmark. IEEE Trans. Cybern. **52**(8), 8114–8127 (2021)
22. Song, S., Lichtenberg, S.P., Xiao, J.: Sun RGB-D: a RGB-D scene understanding benchmark suite. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 567–576 (2015)
23. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1746–1754 (2017)
24. Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E.: Multi-view convolutional neural networks for 3D shape recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 945–953 (2015)
25. Wald, J., Dhamo, H., Navab, N., Tombari, F.: Learning 3D semantic scene graphs from 3D indoor reconstructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3961–3970 (2020)
26. Wang, K., Yin, Q., Wang, W., Wu, S., Wang, L.: A comprehensive survey on cross-modal retrieval. arXiv preprint arXiv:1607.06215 (2016)
27. Wu, Z., et al.: 3D ShapeNets: a deep representation for volumetric shapes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1912–1920 (2015)
28. Xu, Y., Tong, X., Stilla, U.: Voxel-based representation of 3D point clouds: methods, applications, and its potential use in the construction industry. Autom. Constr. **126**, 103675 (2021)
29. Yi, K., et al.: CLEVRER: collision events for video representation and reasoning. arXiv preprint arXiv:1910.01442 (2019)
30. Yuan, J., et al.: SHREC'19 Track: extended 2D scene sketch-based 3D scene retrieval. Training **18**, 70 (2019)