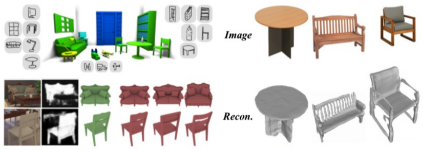
R.Figure 1

R#3: **Silhouette or Saliency?** (object occlusions)



R#4: **SBSR** vs. **IBSR** *R#4:* **3D Reconstruction Data**

ViewPoint	Front	Back
Pix3D	79.5%	75.3%
3D-FUTURE	69.4%	69.4%

R#1: Viewpoints. We evaluate our trained models on the Front (5 views) and Back (5 views) viewpoints. We remove the east and west views. We find our trained models are not sensitive to the azimuths, though 70% of images are captured under front viewpoints.

Online images. Unpreferable elevations will decrease the performance.

Our methods show robustness on most cases.

– ***To R#3 and R#4***. The reviewers **might misunderstand** our motivation and primary contribution when evaluating the paper. The comments such as "Taking efforts to bring SOTA results (R#3)" and "introducing practical applications based on popular adversarial tricks for IBSR (#4)" are the inaccurate understandings of our motivation and idea.

For IBSR, the main challenge is the large appearance gap between 3D shapes and 2D images. The common routine is to map 2D images and 3D shapes into an embedding space and learn a shape similarity measure. We **find that** the shape difference between a negative pair is entangled with the texture gap, making previous metric learning based methods ineffective in pushing away negative pairs. To tackle the issue, we **propose** an elegant idea (as stated by R#1 and R#2) to create hard triplets via the exploration of texture synthesis based on the properties of the IBSR subject. The generation of a hard triplet is shown in Fig. 1, **where** the positive example and the anchor are identical in shape (geometric details), but differ in texture; the negative example and the anchor are similar in texture, but differ in shape. We generate hard triplets in an online manner to improve the cross-domain shape similarity learning. Our **second contribution** is to introduce the saliency attention and viewpoint guidance mechanisms to remedy clutter background noises and unconstrained viewpoints issues of 2D nature images. **Both our motivation and idea are novel** and have been intensively studied in the experimental section. We believe our clever idea, *i.e.,* **generating texture to suppress the adverse impacts of texture**, is potent for future IBSR studies. Besides, our method ranks **1st** on the 3D-FUTURE AI Challenge (IBSR Track) (28 submissions on the leaderboard).

– ***Connection to the NeurIPS community (R#3 & R#4)***. **Firstly**, NeurIPS accepts subjects such as Computer Vision, Application, Information Retrieval, and Embedding Approaches. **Secondly**, IBSR is a fundamental subject in 3D Vision, and the community has put many efforts in building 3D datasets to support the studies of IBSR [1,2,3,4,59,60]. **Thirdly**, with the growing number of 3D shapes, the studies of IBSR is significant. For example, it can help to build 3D virtual scenes for real-world houses by accurately identify the exact 3D shapes contained in the captured 2D scene images. High-performing IBSR systems may also inspire and benefit studies of 3D object reconstruction from large-scale shape collections. Regarding its significance, the retrieval accuracy of IBSR is not that promising than the counterpart, *i.e.*, image retrieval. We thus believe our work in this paper benefits to the advancement of the research on the subject.

– ***Misunderstandings (R#3 & R#4)***. "The major claim of this paper is that by **rendering hard triplets** (**R#3**)": We **have not** claimed to "render" hard triplets. In fact, we can not render hard triplets since each 3D model contains a single UV atlas (texture) in the datasets. 90% of models in ShapeNet [35] are without texture. "**Silhouettes is better than Saliency** (**R#3**)": See R.Fig. 1, we care about the sofa instead of the coffee table. Since both the objects have their own "Silhouettes", "Saliency" is a better choice. There are lots of occluded objects [1,2]. "I would not use the word '**attention**' for the viewpoint module, as it is **simply a weighted sum.** (**R#3**)": Firstly, we use "guidance" instead of "attention". Secondly, most of "attention" strategies can be concluded as weighted sum operations. "**The paper only compares performance with IBSR methods.** (**R#4**)": Firstly, We study IBSR instead of sketch-based shape retrieval (SBSR). Secondly, IBSR focus more on the fine-grained geometric differences, while SBSR retrieves roughly similar shapes in category level. "**Explain the differences between the input datasets and 3D reconstruction datasets.** (**R#4**)": These datasets can be used for 3D reconstruction studies. However, researchers prefer to use rendered synthetic images (without backgrounds and occlusions) based on ShapeNet [35] for the 3D reconstruction subject.

– ***Syntactical corrections and suggested additional explanation for the saliency attention module (SAM) (R#1)***. Thanks for the constructive suggestions. We will correct these syntactical errors. For the SAM, the feature maps are the side features of the ResNet blocks ($\text{Conv}_{4x} \sim \text{Conv}_{32x}$) as presented in Supp-Tab. 2. We will make it clear in the paper.

– ***More discussions for the Viewpoint Guidance (VG) mechanism (R#1)***. For the indoor benchmark (Pix3D [2] and 3D-FUTURE [1]), we find that 70.3% of images are under the front viewpoints. The elevations mainly range from 20 to 45 degrees. The viewpoint distribution makes sense since most people prefer to captured indoor images from front views. For the outdoor Car benchmarks [3,4], the azimuth distribution follows a uniform distribution. We only study azimuths to make consistency with multi-view 3D representations. Our final version will address all the suggestions.

– ***Suggest to make the "particular properties" clear (R#2)***. Thanks for the constructive suggestion. The particular properties are: (1) In contrast to category-driven metric learning for image retrieval, each 3D shape is an individual instance (category). It's not easy to apply hard sample mining strategies in image retrieval to IBSR. (2) 2D objects with different appearances (in practice) may correspond to one 3D shape. (3) While appearance information is a strong feature for 2D image understanding, it has adverse impacts for IBSR. IBSR cares about the geometric details. Based on the properties and our observations, we believe our idea may be a potent venue for future high-performing IBSR studies.