Supplemental
Fine-Tuned but Zero-Shot 3D Shape Sketch View Similarity and Retrieval

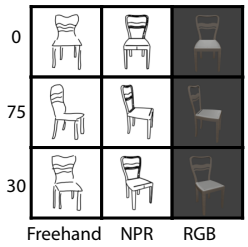
Gianluca Berardi[1,2] Yulia Gryaditskaya[2]

[1]Department of Computer Science and Engineering (DISI), University of Bologna, Italy
[2]CVSSP and Surrey Institute for People-Centred AI, University of Surrey, UK

1. How well can CLIP differentiate between views and objects?

Our work targets retrieval from freehand quick and abstract sketches, such as sketches collected by Qi et al. [2] and Zhang et al. [5]. There are several strategies to represent 3D shapes using their multi-view projections in sketch-based 3D model retrieval literature: using RGB renderings or NPRs (Non-Photorealistic Renderings). However, there is a large domain gap between such views and freehand sketches we consider as queries. To motivate our work, we study CLIP [3] features similarity across domains, and its ability to match the model views across various domains.

In this study, we only consider 3D shapes from the chair category from the *AmateurSketch* dataset [2]. We use three different views for every object, with the camera azimuth angles set to 0°, 30°, and 75°. In this study, we compare shape representation in three styles: (1) *freehand sketches* – sketches from [2] created by human participants without art training; (2) *NPRs* – synthetic sketches created using object-space non-photorealistic rendering with the use of Blender Freestyle [1] or (3) *RGB renderings*. The inset shows example viewpoints in the considered styles for one of the considered 3D shapes.



Further, in this study, we use the third layer of the pre-trained ResNet101 CLIP image encoder as our CLIP embedding space, following [4].

1.1. Study 1: Aggregated feature similarity

First, we consider 10 randomly-selected objects (the corresponding freehand sketches for them are shown in Fig. 1), and compute the distance between two 3D shapes A and B, represented with their multi-view renderings or freehand sketches from the three considered viewpoints, as follows:

$$d(A^i, B^j) = \frac{1}{V}\sum_{k=1}^{V}\Big(\mathrm{CLIP}(A^i_k) - \mathrm{CLIP}(B^j_k)\Big)^2 \quad (1)$$

where $V = 3$ is the number of views, and subscripts i and j denote one of three image domains: freehand sketches, NPR renderings, or RGB renderings. Since in this study we *aggregate* the distance across matching viewpoints in different domains, we refer to this study as *'Aggregated feature similarity'*.

In Fig. 1, we plot pairwise distances between shapes, when their views come from one of the three image domains. We can see that in all three configurations, comparing the same object between different domains results in the lowest average distance (darker color) in most of the cases. *This shows the **general robustness of the CLIP model** across different domains. We also can see that it is easier to find correct matches for 3D shapes represented with freehand sketches and NPR renderings than for 3D shapes represented with freehand sketches and RGB renderings.*

1.2. Study 2: Matching views across various domains

Comparing individual views k and h in domains i and j, we compare view embeddings of the same object for the compared domains, and average over different objects:

$$d(k^i, h^j) = \frac{1}{N}\sum_{A=\mathrm{obj}_1}^{\mathrm{obj}_N}\Big(\mathrm{CLIP}(A^i_k) - \mathrm{CLIP}(A^j_h)\Big)^2 \quad (2)$$

where $N = 100$ is the number of objects. Fig. 2 shows that for all configurations the average lowest values are obtained when comparing the same view. *This shows that, in general, CLIP is able to match the same view of an object between different domains. Similarly, we observe that matching views of freehand sketches with NPR renderings is more*

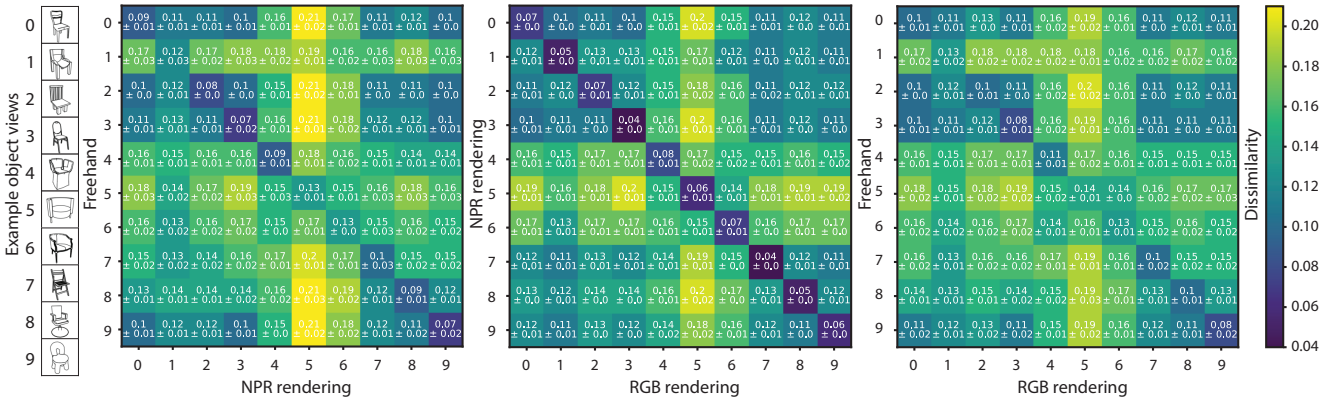


Figure 1. We plot pairwise distances between shapes when their views come from one of the three image domains: freehand sketches [2], line or RGB renderings. Note that in this study, we use geometry-based NPRs (Non-Photorealistic Renderings) [1]. See Sec. 1 for the details.

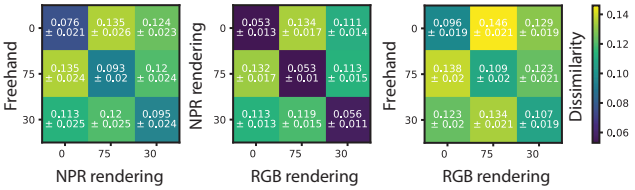


Figure 2. We plot average pairwise distances between views in different image domains: freehand sketches [2], NPR or RGB renderings. See Sec. 1 for the details.

reliable than matching freehand sketches with RGB renderings. However, when comparing freehand sketch views with views from other domains, we observe that the confidence intervals can overlap. This means, that ***occasionally an incorrect viewpoint in a different domain can be selected****.*

1.3. Motivation for our work

These two studies motivate our work on considering diverse fine-tuning strategies. In particular, we challenge ourselves with the task where fine-tuning can be performed on a single shape category and study the model's generalization to other categories. We refer to the latter as *'Fine-Tuned but Zero-Shot'*.

References

[1] Blender Freestyle: https://docs.blender.org/manual/en/latest/render/freestyle/introduction.html. 1, 2

[2] Anran Qi, Yulia Gryaditskaya, Jifei Song, Yongxin Yang, Yonggang Qi, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Toward fine-grained sketch-based 3d shape retrieval. *IEEE Transactions on Image Processing*, 30, 2021. 1, 2

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 2021. 1

[4] Yael Vinker, Ehsan Pajouheshgar, Jessica Y. Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. Clipasso: Semantically-aware object sketching. *ACM Trans. Graph.*, 41(4), 2022. 1

[5] Song-Hai Zhang, Yuan-Chen Guo, and Qing-Wen Gu. Sketch2model: View-aware 3d modeling from single free-hand sketches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1