Zero-Shot Multi-Modal Artist-Controlled Retrieval and Exploration of 3D Object Sets

Kristofer Schlachter*
kristofer@unity3d.com
Unity Technologies
U.S.A.

Benjamin Ahlbrand*
ben.ahlbrand@unity3d.com
Unity Technologies
U.S.A.

Zhu Wang
zhu.wang@nyu.edu
New York University
U.S.A

Valerio Ortenzi
valerio.ortenzi@unity3d.com
Unity Technologies
Denmark

Ken Perlin
perlin@cs.nyu.edu
New York University
U.S.A

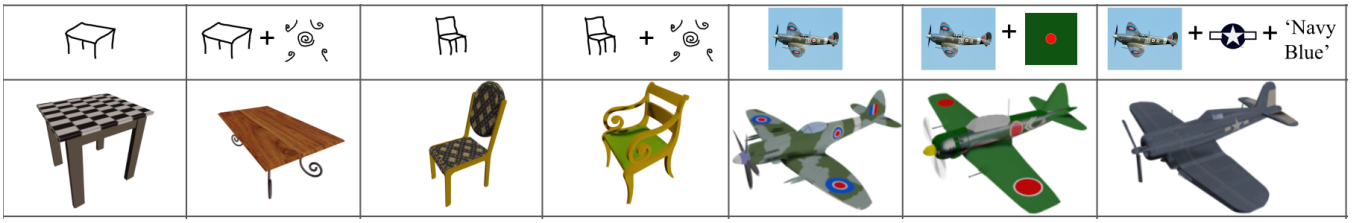

Figure 1: Examples of retrieval with only the top match shown. Top Row: Queries containing sketch, image, and text inputs to the 3D mesh retrieval system (weights are not shown in the multiple input queries). **Bottom Row:** Top match in the 3D mesh database based on the similarity score of the query's embedding with the embedding of the renderings of the 3D meshes.

ABSTRACT

When creating 3D content, highly specialized skills are generally needed to design and generate models of objects and other assets by hand. We address this problem through high-quality 3D asset retrieval from multi-modal inputs, including 2D sketches, images and text. We use CLIP as it provides a bridge to higher-level latent features. We use these features to perform a multi-modality fusion to address the lack of artistic control that affects common data-driven approaches. Our approach allows for multi-modal conditional feature-driven retrieval through a 3D asset database, by utilizing a combination of input latent embeddings. We explore the effects of different combinations of feature embeddings across different input types and weighting methods.

CCS CONCEPTS

• **Computing methodologies** → **Graphics systems and interfaces**; **Machine learning**; **Search methodologies**; **Perception**.

KEYWORDS

neural networks, computer graphics

ACM Reference Format:
Kristofer Schlachter, Benjamin Ahlbrand, Zhu Wang, Valerio Ortenzi, and Ken Perlin. 2022. Zero-Shot Multi-Modal Artist-Controlled Retrieval and Exploration of 3D Object Sets. In *SIGGRAPH Asia 2022 Technical Communications (SA '22 Technical Communications), December 6–9, 2022, Daegu, Republic of Korea.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3550340.3564216

1 INTRODUCTION

The creation of 3D content generally presents a high barrier to entry, since it requires specialised skill sets even when using existing modelling software. While artists and other experienced content creators can navigate their way through labor-intensive pipelines, other less skilled users are often discouraged and tend to give up. For example, hobbyist world builders are generally able to produce a rough sketch of a 3D model, but then often find the process of creating the corresponding 3D content to be long and difficult. Furthermore, we observe a disconnect between creative workflows and existing sketch-to-mesh retrieval-based methods, as such methods tend to lack the artistic control needed to explore the space of results.

In this paper, we address the problem of democratising content creation by enabling 3D asset retrieval from 2D images, simple 2D sketches, and text input. In particular, our approach allows users to define the weights of inputs by leveraging CLIP (Contrastive Language-Image Pre-Training) embeddings[Radford et al. 2021]. We strongly believe that a system which accepts diverse inputs such as sketches, 2D images and text can further lower the barrier

*Both authors contributed equally to this research.

SA '22 Technical Communications, December 6–9, 2022, Daegu, Republic of Korea

ACM ISBN 978-1-4503-9465-9/22/12…$15.00
https://doi.org/10.1145/3550340.3564216

to entry for 3D content creation. From this perspective, the main contributions of this paper are:

- A novel pipeline with zero-shot sketch-based multi-modal retrieval.
- A weighted interpolation in the latent space of (visual style independent) multi-modal inputs to enable artistic control over the retrieval sets.

2 RELATED WORK

Our work builds upon recent CLIP-like models. CLIP is a neural network optimized over 400 million natural images and matching text descriptions. Its embedding space allows "semantic alignment" of features between images and text passed as input. Our proposed method enables a feature extraction which allows manipulation of the latent space such that desirable semantic traits can be retrieved without the need for a complex dataset or training process.

2.1 CLIP-driven Generative Models

Other works use CLIP-encoded 2D images and text to accomplish tasks involving 3D meshes. Text2Mesh[Michel et al. 2021] and Text-to-Mesh[Khalid et al. 2022] optimize 3D deformations of 3D meshes by combining a differentiable renderer with supervisory CLIP text prompts. Text2Mesh also explored optimizing towards multiple input targets by summing the embeddings. Our methods allow for arbitrary weightings of embeddings and real-time feedback of such effects.

We use 2D CLIP encoded views of a 3D object in a similar manner to Dreamfields[Jain et al. 2022]. However, in [Jain et al. 2022] the similarity of the input text prompt supervises the optimization across multiple camera views while our approach uses these prompt features to compute similarity for retrieval.

Our work uses CLIP to semantically understand drawings. This was also done in CLIPASSO[Vinker et al. 2022] where the authors used it to ensure increasingly abstract sketches that are semantically similar to the supervisory text prompt. Concurrent work to ours called TASK-former[Sangkloy et al. 2022] sums the embedding of a single sketch and a single text prompt to retrieve images. That work shows how using a text prompt embedded by CLIP can achieve state-of-the-art results via text-based image retrieval. Our work differs by allowing any number of inputs and modes (in fact, text is not required), and allowing custom weights on those inputs; furthermore, we do not need to train on a dataset, due to our zero-shot approach.

2.2 Zero-Shot Sketch-based Retrieval

There are a number of recent works on zero-shot sketch-based image retrieval. Domain Disentangled GANs (DD-GAN)[Xu et al. 2022] train a generative adversarial network on SHREC '13[Li et al. 2013] + SHREC '14[Li et al. 2014] in order to build a mapping between user-based sketches and the 3D model to be retrieved. In practice, this is a brittle approach, since a wide range of inputs can be objectively "semantically correct" in representing a shape.

Tursun et. al.[Tursun et al. 2022] use a pre-trained model as a teacher to help learn abstract labels for their zero-shot sketch-based image retrieval (SBIR) model. Dutta and Akata[Dutta and Akata 2019] use an adversarial loss and enforce cycle-consistency on the shared latent space to close the domain gap. In comparison, CLIP uses contrastive learning and text descriptions of 29 times more images than ImageNet[Deng et al. 2009] to create a shared latent space.

To our knowledge, all existing sketch-based retrieval methods (including zero-shot) return a one-to-one mapping between sketch and result. They are further limited by a single sketch style and by a lack of multi-modal input, as well as by a single representation (style, level of abstraction, etc).

Previous works[Michel et al. 2021] have also shown that 2D viewpoint supervision has been effective for 3D tasks. Those same works have emphasized that changing the viewpoint of the object doesn't change its semantics. In contrast, our work allows the user to distort the mapping through a combination of many inputs and/or weighting of those inputs in order to explore the output space. In this work, we explore the effects of different combinations of inputs across different media types and weighting methods.

3 METHOD

3.1 Zero-shot learning using CLIP

Zero-shot learning methods allow for the classification/interpretation of data that was not seen during training. Zero-Shot models produce a rich embedding space that aims to optimize identification of semantically meaningful features, which can then be used to cluster around desired categorizations. This is done through auxiliary information, which in CLIP's case is textual descriptions of images. The learned semantic features are useful for down-stream tasks such as retrieval where unseen images can be embedded into the rich feature space and a distance measure can be performed. We used a pre-trained CLIP model[Radford et al. 2021] because it was optimized on the task of "representation learning". CLIP also has the added benefit of aligning the embedding of multi-modal inputs (in this case images and text) into a common feature-rich embedding space that can be used for zero-shot based retrieval.

3.2 Data Setup

Retrieval of 3D models is enabled by rendering each model from three different viewpoints (front, back and perspective) and storing the corresponding images along with each mesh. Then each image is encoded through CLIP and that embedding is stored along with the model. We found that these three views lead to valid matches for the specific models that were in the dataset.

The dataset itself is 5 categories of ShapeNet[Chang et al. 2015]: sofa, chair, table, car and plane (roughly 25K meshes). We embedded them using three different rendering setups: the original ShapeNet 3D mesh, the original model without textures, and finally a decimated, smoothed and untextured mesh. The untextured configurations helped explore the effect of texture and detail on matching for retrieval.

3.3 Multi-Modal Embedding Fusion

These 2D renderings allow retrieval on 3D models to be effective. Since CLIP has aligned the embedding between text and images, queries for retrieval can take the form of text prompts and images. Sketches are passed through the CLIP image encoder and can be combined with an arbitrary number of other text or image inputs.

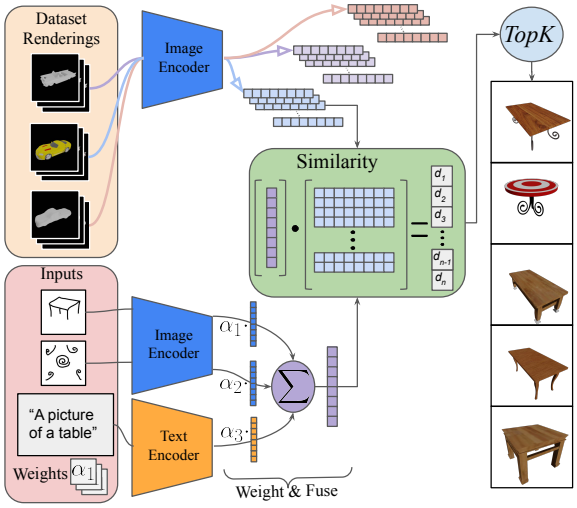

Figure 2: Architectural overview of our method. The dataset is rendered three ways (textured, untextured, untextured and smoothed). The images are then embedded using CLIP's image encoder. Each query input is separately CLIP-encoded, weighted and then fused into a single embedding. The fused query is then used to score each rendering in the active dataset (in this example textured) and the top 5 similarities per mesh are retrieved.

Each is encoded and then combined into a single 512 dimension normalized code. After computing the cosine similarity score (see Equation 2) between that code and every 3D mesh's rendered images, we sort and filter (thereby removing duplicate matches that may result from multiple views being in the database) the top-k nearest neighbors. See Figure 2 for more details on the data flow.

Embedding code fusion takes place in prior work[Michel et al. 2021] through summing and averaging. We demonstrate this alongside arbitrary weighting for the embeddings. In order to arbitrarily weight inputs, we expose a single float per input that we then multiply by before summing (as opposed to simply summing unmodified embeddings).

This weighting enables artistic control, in order to allow interactive refinement of the query embedding and to direct the user's search as a natural process. We show our weighted interpolation of multiple embeddings in Equation 1, where n refers to the number of inputs, z refers to the embedding, and α is the user specified weights.

$$fusion = \Sigma_{i=1}^{n} \alpha z \tag{1}$$

$$similarity = [query][features]^T \tag{2}$$

4 RESULTS AND DISCUSSION

To illustrate the flexibility of our method, we explore different image inputs, as well as the effect of texturing on retrieval queries containing high frequency class independent features, in addition to enabling fine-tuning different weighted combinations to arrive at a result.

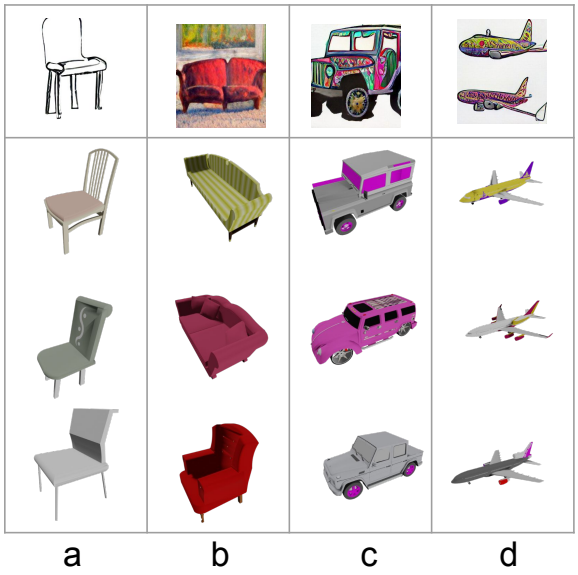

Figure 3: Examples of different input sketch styles and details. Top Row: Inputs representing different level of abstraction and style. **Other Rows:** The top matches to the input query where the closest match is the top row and the next ranking ones are below.

4.1 Robust Semantic Encoding of Images

To demonstrate the robustness of matching through the style-invariant semantic embedding of input sketches, we chose a variety of artistic visual styles that were distinct from the style of the rendered objects in the dataset. Figure 3 shows that a varied range of image fidelity, drawing methods and styles can produce valid matches. Retrieval is also able to handle incomplete renditions of objects, as in the chair and truck queries. Semantic understanding of the query is also not impacted when multiple objects are in the image as is the case for the couch and airplanes. This may be due to the use of a zero-shot learner where features are learned over category labels with high level features to allow a correct match at various levels of abstraction.

4.2 Fine Details Inform Retrieval

Figure 3 also shows that more detailed sketches can lead to more specific matches. Adding color, texture and details influence which model is retrieved. These cues, when available, disambiguate objects that belong to the same semantic category. This can be seen in the matching of the pink texture of the truck and more interestingly the results of the two airplane query matches airplanes that incorporate coloring of both. A shortcoming of the color matching is that the top scoring couch is actually the color of the background. It is the second ranking match which has the color of the couch in the query image.

4.3 Multi-Modal Input and Refinement

As we can see, steering a high level semantic match through a single sketch can be done by altering the strokes or by adjusting color and texture. However, this way of querying requires more artistic skill to accurately express one's intention. Instead, using a different search paradigm that combines multiple targets into the query allows for

expressing a search concept without requiring the skill to create an accurate visual representation. By leveraging the learned semantic features of the CLIP encodings and fusing multiple inputs that are feature specific instead of object class specific, an intuitive form of semantic based retrieval can be explored. Figure 1 shows how this can be accomplished by providing a simple high level sketch and auxiliary examples of stylistic flourishes to furniture, insignias on an aircraft, or text describing paint color.

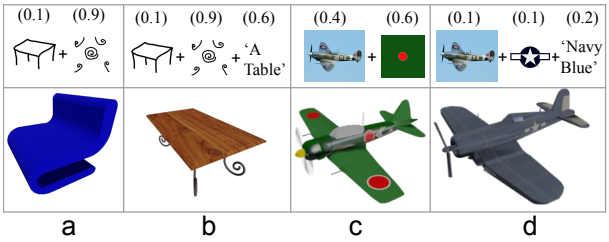

Figure 4: Examples weights shown for sub-feature matching. Top Row: Inputs. The weights of the inputs are in parentheses and are not required to sum to one. **Bottom Row:** Top matching result. In column 'a', notice the failure to match a table; adding a text prompt to the query improved matching. Columns 'a' and 'b' are matched against the untextured dataset to enforce geometric feature matching. The columns 'c' and 'd' were matched against the original textured models because the examples included textures.

4.4 Limitations and Steering The Result Set

Figure 4(a) shows that the query's encoding matches a chair better than a table. It is also clear that a large flat surface and curves have been recognized and used for the match but are combined in an unintended way. A solution is to reduce the space of possible matches: Figure 4(b) incorporates a text prompt that only encodes the features of a table and compels matches to be from the set of tables in the dataset.

Other failure cases show up as interpreting the star insignia not as a texture but as geometry and matching furniture views that have the same rough outline. The context of the features also matters, just a red circle as an auxiliary input did not succeed in finding an airplane with that insignia, scaling it down and adding a green background resulted in the match shown in Figure 4(c). There is also a bias in the matching towards a subset of meshes if no auxiliary inputs are given. Conditioning on other inputs compensates for the biases of the encodings and allows wider exploration of the set of meshes. Some of the linear combinations of weights may not return intended results. Crucially, instant feedback to weight changes (less than 10 milliseconds on Titan RTX to query set of more than 100K embeddings) allows exploration of the fusion-space in real-time and thereby provides a practical way to interactively refine the result set to fit with one's artistic intentions.

5 FUTURE WORK

Future work should include exploring other text + image multi-modal models aside from CLIP such as DALLE-2[Ramesh et al. 2022], and Imagen[Saharia et al. 2022], as well as fine tuning such models to provide better results on zero-shot sketch retrieval tasks. We leave benchmarking as future work as we were unable to find any that necessarily included multi-modal inputs for zero-shot retrieval. Our contribution to combining latent features is also applicable to other use cases such as image generation.

6 CONCLUSION

In this paper we addressed the problem of high-quality 3D asset retrieval from multi-modal inputs, including 2D sketches, images and text. We used CLIP features to perform multi-modality fusion in order to address the lack of artistic control that affects common data-driven approaches. By aligning text and images we were able to perform retrieval at high levels of abstraction. By using embeddings of 2D views of 3D meshes for retrieval we successfully captured view independent properties and semantics of the mesh, and allowed for effective retrieval. Finally, our weighted multi-input fusion created a pathway to help mitigate the biases and the idiosyncrasies of CLIP's embedded space.

REFERENCES

Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. 2015. *ShapeNet: An Information-Rich 3D Model Repository.* Technical Report arXiv:1512.03012 [cs.GR]. Stanford University — Princeton University — Toyota Technological Institute at Chicago.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.

Anjan Dutta and Zeynep Akata. 2019. Semantically Tied Paired Cycle Consistency for Zero-Shot Sketch-based Image Retrieval. In *CVPR*.

Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. 2022. Zero-Shot Text-Guided Object Generation with Dream Fields. (2022).

Nasir Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. 2022. Text to Mesh Without 3D Supervision Using Limit Subdivision. *arXiv preprint arXiv:2203.13333* (2022).

Bo Li, Yijuan Lu, Afzal Godil, Tobias Schreck, Masaki Aono, Henry Johan, Jose M Saavedra, and Shoki Tashiro. 2013. *SHREC'13 track: large scale sketch-based 3D shape retrieval.*

Bo Li, Yijuan Lu, Chunyuan Li, Afzal Godil, Tobias Schreck, Masaki Aono, Qiang Chen, Nihad Karim Chowdhury, Bin Fang, Takahiko Furuya, et al. 2014. SHREC'14 track: Large scale comprehensive 3D shape retrieval. In *Eurographics Workshop on 3D Object Retrieval*, Vol. 2. .

Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. 2021. Text2Mesh: Text-Driven Neural Stylization for Meshes. *arXiv preprint arXiv:2112.03221* (2021).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv preprint arXiv:2205.11487* (2022).

Patsorn Sangkloy, Wittawat Jitkrittum, Diyi Yang, and James Hays. 2022. A Sketch is Worth a Thousand Words: Image Retrieval with Text and Sketch. *European Conference on Computer Vision, ECCV* (2022).

Osman Tursun, Simon Denman, Sridha Sridharan, Ethan Goan, and Clinton Fookes. 2022. An efficient framework for zero-shot sketch-based image retrieval. *Pattern Recognition* 126 (2022), 108528.

Yael Vinker, Ehsan Pajouheshgar, Jessica Y. Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. 2022. CLIPasso: Semantically-Aware Object Sketching. arXiv:2202.05822 [cs.GR]

Rui Xu, Zongyan Han, Le Hui, Jianjun Qian, and Jin Xie. 2022. Domain Disentangled Generative Adversarial Network for Zero-Shot Sketch-Based 3D Shape Retrieval. *arXiv preprint arXiv:2202.11948* (2022).