ULIP-2: Towards Scalable Multimodal Pre-training for 3D Understanding

Supplementary Material

A. Appendix

A.1. Ablation on 3D Input

In order to fairly compare to OpenShape on Objaverse-LVIS benchmark, which utilizes 10k colored point clouds as the 3D input, we adopt the same 3D input preprocessing as in OpenShape. We conduct this ablation study to assess how color information influences zero-shot classification results on Objaverse-LVIS. To do this, we evaluated the ULIP-2 pre-trained Point-BERT model, pre-trained on Objaverse + ShapeNet. Our findings from Table 10 show that ULIP-2 maintains strong performance on Objaverse-LVIS zero-shot classification tasks, even without using color information.

3D Encoder Input	Objaverse-LVIS	
	top-1	top-5
8k xyz	48.9	77.1
10k xyzrgb	50.6	79.1

Table 10. Point-BERT w/ ULIP-2 zero-shot 3D classification on Objaverse-LVIS, pre-trained on Obajverse and ShapeNet jointly with OpenCLIP ViT-G encoders.

A.2. Different Kinds of 3D Backbones

To verify ULIP-2's improvement is agnostic to 3D backbones, we conduct experiments on the PointNeXt backbone. Results in Table 11 show that, with another kind of intrinsically different 3D backbone, ULIP-2 can still improve the performance significantly. Given that Point-BERT is a scale-up-friendly transformer-based architecture and has better zero-shot classification results, we mainly conduct experiments on Point-BERT for all our experiments.

Model	Pre-train method	ModelNet40	
		top-1	top-5
PointNeXt [36]	ULIP [52]	56.2	77.0
	ULIP-2	72.8	95.7
Point-BERT [55]	ULIP [52]	60.4	84.0
	ULIP-2	**75.2**	**95.0**

Table 11. Zero-shot 3D classification on ModelNet40 with different 3D backbones, pre-trained on ShapeNet with SLIP ViT-B encoders.

A.3. More Details of the Generated Triplets

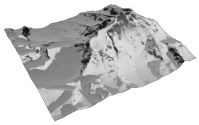
"a cell phone on a tripod with a grey background",
"a camera, tripod and black background on a foggy day",
"a photo camera is on a tripod next to a light",
"a gray photo of a smartphone with a tripod in the background",
"a cell phone on a tripod against a grey background"

"a camera and tripod sitting on the ground in a gray room",
"a man standing at a table with a tripod",
"a picture of a camera that is on a tripod",
"there is a standing man with an unknown device in his hands",
"a white camera sitting on a tripod next to some concrete",

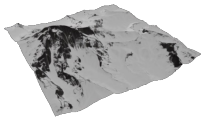
zdYQzx9xxVoNb8OJbR8iQf6GOK6

"a black background with a tripod in the middle of it",
"a picture of an older camera with a tripod",
"a small camera with an orange filter with a tripod",
"a photo of an old camera standing on a tripod",
"an old camera on a tripod sitting in a dark room",

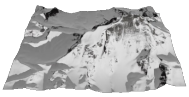
dd826f1f05c544ccb671949ebca59721



"a 3d model of a mountain with a grey background",
"3d mountain texture",
"a mountainside with a grey background",
"some 3d terrain that is very abstract",
"mountains on a grayish background"



"a snowy mountain on gray surface",
"an ice - formed object with the title, \"mountain snow\"",
"3d mountain landscape with a grey backdrop",
"a white piece of paper with a mountain image on it",
"a mountain is on the top of a large paper",



"a mountain in a computer with some snow on it",
"a piece of paper sitting on top of a ground surface",
"a picture of a plain white sheet on a gray background",
"a photo of a piece of paper",
"a folded paper wrapped around some type of object",

ea79b3b36fc04ded845ddfe1de08957a

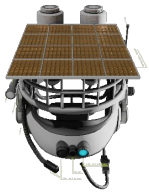
"3d objects with lots of trees around them",
"the rendering shows the large area of land to be destroyed",
"the front of a big ship on the horizon",
"3d model of a tall bridge with multiple tall trees around the top of the bridge",
"a rendering of some futuristic objects on a gray background",

"a gray background of a white thing sitting",
"a picture of some buildings in an empty background",
"a brown image of some objects that are shaped like buildings",
"a picture of the side of a building",
"an image that looks like it was made out of snow",

"a gray background of a white thing sitting",
"a picture of some buildings in an empty background",
"a brown image of some objects that are shaped like buildings",
"a picture of the side of a building",
"an image that looks like it was made out of snow",

x7dkxRTdn5yn5tmlS1umdwzM4in

"a computer model showing two items attached to one another",
"a model of a device with two large cylinders",
"an airplane has a lot of boxes on top of it",
"a computer rendered picture shows a tower with some tanks on top of it",
"a detailed representation of an electrical generating device",



"a 3d model of a robot's solar panel and water tank",
"a small drawing showing the basic construction of a solar panel",
"an unmanned machine with solar panels on top",
"a diagram of a solar powered boat showing a computer-generated diagram",
"a picture of a device that is solar powered",

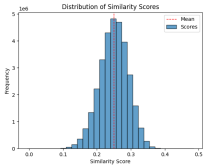
"a cart with various parts that includes a red cross symbol",
"a diagram of an object that has a medical device on it",
"a wooden structure is shown with medical equipment",
"a diagram of the setup for a medical treatment plan",
"the electronic system shows how an automated medical system functions",

ff7941803f9e490c9847cac2b85c48c4

"a spinning round shape is shown on a grey background",
"a drawing of a white ring that is in an image",
"a close up of a circular object on a greyish background",
"a picture of a fire pit with a white background",
"a small white object with a ring around it on a dark gray background",

"a circular object with something in the middle",
"the metal structure is spinning on the floor",
"a very cool looking computerized object",
"the inside of a round object is on a dark background",
"a white object with a round bottom is shown",

"a big circular object is in the middle of a blurry gray background",
"a picture of a ship wheel on a gray background",
"a white disc sitting in a darkened room",
"a picture of a steering wheel on a grey background",
"an abstract piece of art that looks like a circle",



The CLIP similarity distribution for "ULIP-ShapeNet" Triplets.

Figure 4. More qualitative samples from ULIP-Objaverse Triplets.