Supplementary Material for Intrinsic Object-Centric Image Similarity

Klemen Kotar*, **Stephen Tian***, **Hong-Xing Yu**, **Daniel L. K. Yamins**, **Jiajun Wu**
Stanford University
*Equal contribution
{klemenk, stephentian, koven, yamins, jiajunw}@stanford.edu

A Dataset Details

Our dataset consists of a total of 50 categories with 2 or 10 instances each, for a total of 180 objects. The specific categories are listed below:

1. Apple
2. Strawberry
3. Orange
4. Pear
5. Apricot
6. Banana
7. Mango
8. Broccoli
9. Carrot
10. Potato
11. Yellow onion
12. Shallot
13. Tomato
14. Grapes
15. Lettuce
16. Avocado
17. Bell pepper (yellow)
18. Bell pepper (orange)
19. Bell pepper (red)
20. Can of tomatoes
21. Eggplant
22. Green water bottle
23. Marker
24. Pencil (blue)
25. Pen
26. Mug
27. Ceramic Mug
28. Fork
29. Spoon
30. Butter Knife
31. Spatula (wood)
32. Cups
33. Ceramic pot
34. Plate
35. Bowl
36. Notebook
37. Book
38. Diet Coke
39. Red soda
40. Rose
41. Cookie
42. Octopus
43. Cardboard box
44. Mouse
45. Keyboard
46. Cable
47. Screwdriver
48. Pliers
49. Hammer
50. Screw

Additionally, ten categories contain ten object instances. These categories are enumerated below:

1. Apple
2. Banana
3. Pen
4. Ceramic Mug
5. Fork
6. Spoon
7. Book
8. Mouse
9. Screw
10. Screwdriver

Submitted to the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks. Do not distribute.

Light setting	Shutter speed	F-number	ISO	Focal length
Left, back, right	$1/640$s	f/2.8	2000	39mm
Low light	$1/20$s	f/2.8	2000	39mm

Table 1: Camera settings for our dataset.

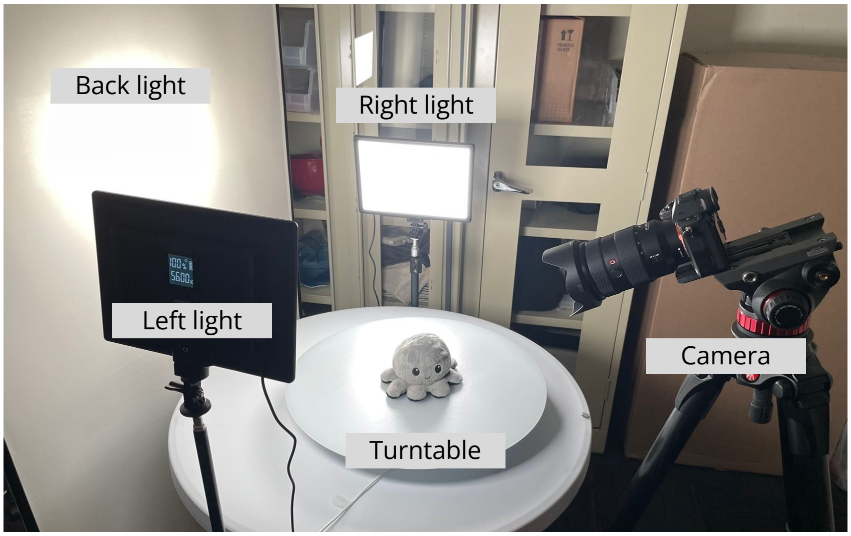

Figure A1: Capture setting for our controlled **different illumination** and **different object poses** configurations.

We capture images using a Sony α 7IV mirrorless camera equipped with a FE 24-70mm F2.8 GM lens. The camera settings used in our **different illumination** and **different object poses** data capture configurations are enumerated in Table 1. We show our capture setup in Figure A1.

For the **different object poses** data capture configuration, we use the turntable to perform a full 360 degree rotation at a constant velocity with a period of approximately 24 seconds. We then configure the camera to capture 24 images shooting at an interval of 1 second.

We capture images in the controlled setting at a resolution of 3168×3168.

For "in-the-wild" settings, we capture images using cell phones (iPhone 13, iPhone 14 Pro, iPhone 12 Mini) at varying resolutions.

B Data

Dataset description. We provide a dataset description in a dataset sheet: `https://github.com/s-tian/PlatonicDistance/blob/main/datasheet.md`

Link and license. The dataset is uploaded for public download under the CC-BY-4.0 license: `https://purl.stanford.edu/gj714cj0414`.

Maintenance. Our dataset is hosted on the Stanford Research Data digital repository which will provide long-term support for hosting the dataset. It also provides structured metadata (`schema.org` standards) It has the following DOI: `https://doi.org/10.25740/gj714cj0414`.

Author statement. The authors bear all responsibility in case of violation of rights. All dataset images were collected by the authors and we are releasing the dataset under CC-BY-4.0.

Format. The data is uploaded in a simple `zip` format. Upon decompressing the archive, a directory is provided for each object category. In each directory, the data is further split into "instance_1" and "instance_2" which represent the two object instances for each category. Under these directories, JPEG images are stored for each lighting condition, with file name descriptors indicating the relative angle in degrees of rotation under which the object was taken, as well as in-the-wild images, which are labeled only with an arbitrary index 0 through 4.

C Potential societal impacts

Our paper introduces a metric that may find a range of downstream applications, including improving the temporal consistency of generative models, augmenting vehicle and human re-identification, to aiding perception for embodied agents. As with any improvements, e.g., in consistent video generation, nefarious actors may seek to use these technologies for harmful purposes. We recognize that there may be additional future applications that we cannot currently foresee.

We see opportunities for this research and our findings in Section 5.1 to provide some inspiration for additional research in neuroscience on understanding the human medial temporal lobe.

By introducing a new evaluation dataset, we are also providing a benchmark that may impact the direction of future work. For example, our dataset consists of common objects largely found around North American homes and laboratories, and images are captured in outdoor settings in a limited set of geographical locations. While we select items that we believe are commonly occurring around the world, performing future evaluation on just the CUTE dataset may create a bias towards particular types of objects or scenes. Our metric also largely inherits any biases present in the original DINOv2 model.

D Computational resources

The computational resources used were a personal workstation and computing nodes from the Stanford SC computational cluster. We used a personal workstation with an NVIDIA RTX 3090 GPU for the main experiments, and on the SC cluster, we used around 30 jobs lasting at most 2 hours each to perform the Re-ID experiments, including the sweeps on the values of α. We use 1 NVIDIA TITAN RTX GPU for each job. We also ran additional backbone ablations on the SC cluster with around 30 jobs lasting at most 4 hours each using one NVIDIA A40 GPU each.

E Qualitative Analysis

One characteristic of the proposed metric is its continuous nature. Different from recognition tasks such as Re-ID, it does not simply classify two objects as being the same instance but rather measures how similar they are. Since a measure of distance in the space of all objects is highly subjective and a definitive ground truth is exceedingly difficult to establish, we evaluate our metric using human preference.

Our study design is as follows: one of the authors generated 10 sets of 5 images each, consisting of photos taken by the author and photos from the internet. Then 2 other authors each ranked the image sets, considering both personal preference and the demonstration quality of the set. Their votes were averaged out and the top 5 image sets were selected. These image sets were then scored by LPIPS, CLIPScore, and foreground feature averaging (FFA) and ordered from most similar to least similar to a query object in each set. 34 participants then chose which ordering they prefer according to their personal subjective opinion. The specific prompt they were given was: "This is a quick, anonymous survey about ordering preference. Please carefully look at each set of images. 3 Orderings are presented for each set. Select the ordering that makes the most sense to you. The images are ordered from most similar to least similar to the first image (left to right). There is no one correct answer, we seek your subjective opinion."

In Figure A2 we see that FFA was the top choice on 4 out of the 5 sets and a close second choice on the fifth. While the participants showed strong agreement on certain sets, they generally displayed pretty mixed opinions, highlighting the subjective nature of this type of classification. The given prompt was intentionally vague, allowing the participants to focus on various aspects of the images such as the pose of the objects, the background, the class, etc. Despite this, the results of this limited study suggest that our proposed metric is reasonably aligned with the intuitive human definition of similarity.

Figure A2: Human survey results. Each of the five groups of images represents an image set. For each image set, we generated three orderings, based on LPIPS, CLIPScore, and foreground feature averaging (FFA) respectively. For each image set, participants were asked to determine which of the three orderings they preferred, where the ordering represented a ranking of the similarity of the first image to the others. We see that in four out of five cases, participants preferred orderings scored as in FFA.

F Code and instructions for experiments

The code can be found at: `https://github.com/s-tian/PlatonicDistance`.

Experimental details and hyperparameters To strike a balance between performance and speed we chose the DINOv2 ViT-B/14 distilled backbone (dinov2_vitb14) for FFA, consisting of 86M parameters. DINOv2 models are capable of accepting inputs at various resolutions, but we select a fixed input size of 336×336 for the same reason as above, and also to provide a fair comparison to the CLIPScore metric. Our CLIPScore is based on the ViT-L/14@336px model from OpenAI.

In order to obtain the foreground mask we pass the input image through the Tracer-B7 model provided by CarveKit. We then downsample the foreground mask by a factor of 14 (the DINOv2 patch size) in order to obtain a mask of the same size as the DINOv2 feature grid. We then superimpose the two and average the unmasked patches.

G Re-ID Experimental Details

We select the hyperparameter α for the weighted sum between SpCL and each considered model by sweeping over the values $[0.1, 0.2, ..., 0.9]$ on a validation set and picking the α value with the best top-1 accuracy on the validation set. The score is always computed by summing α times the model in question and $1 - \alpha$ times the SpCL score. The selected values of α for each model are reported below:

Model	VeRi	CityScapes
SpCL+LPIPS	0.1	0.1
SpCL+DINOv2+Crop	0.5	0.9
SpCL+DINOv2 (Global)	0.1	0.2
SpCL+FFA DINOv1 (Crop-Img)	0.7	0.9
SpCL+FFA (Crop-Img)	0.6	0.9

Table 2: Selected α values for ReID experiments combining SpCL with various metrics.

H Relation to Other Metrics of Similarity

There are many aspects to object similarity. One can measure the visual similarity - such as the shape, color or texture of objects or functional similarity such as the purpose or affordance of an object. Often these measures are entirely orthogonal to each other, and further influenced by the context of the comparison. Because of this, many measurements of object similarity lack proper ground truth. Our aim is to define one particular dimension of similarity where we can obtain at least partial ground truth labels. We can obtain strong binary labels for this particular metric based on the identity of the objects themselves – different images of the same object should have an ideal perfect similarity.