Benchmarking Omni-Vision Representation through the Lens of Visual Realms

Yuanhan Zhang[1], Zhenfei Yin[2], Jing Shao[2], and Ziwei Liu[1]

[1] S-Lab, Nanyang Technological University, Singapore
{yuanhan002,ziwei.liu}@ntu.edu.sg
[2] SenseTime Research
{yinzhenfei,shaojing}@sensetime.com

1 Annotation Interface

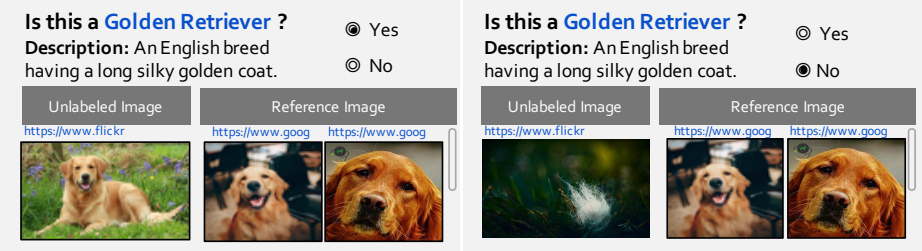

Fig. 1. The Annotation Interface for OmniBenchmark. The meta information of each concept consists of a tag, description, and reference images. Annotators determine whether the unlabeled image conforms to the concept. The yes option will be marked if the annotator provides a positive answer to the question above.

We illustrate the annotation interface for building OmniBenchmark in Fig. 1. In addition, the annotator needs to determine the legality of the image. If the unlabeled images contain pornography or any offensive content, annotators should click "No" button in this interface.

2 Experiments Results

We present the complete Top-1 accuracy (Table 1) and accuracy relative to the baseline ResNet50 (Table 2) of all the 26 models. For all the models, we split them across the four model categories. As shown in Table 1, Swin-T that has similar parameters to ResNet-50 models achieves the best average accuracy on the OmniBenchmark. In addition, GV-D [17] is the best ResNet-50 model, which is pre-trained on the 69M GV-D data.

3 Social Impact

The usage of OmniBenchmark might bring several risks, such as privacy, and problematic content. We discuss these risks and their mitigation strategies as follows.

Copyright. All images in this paper and the dataset are licensed by the CC-BY (https://creativecommons. org/licenses/by/2.0/) license and contain information on their creator. On the above url, this license is described as follows:

- Copy and redistribute the material in any medium or format.
- Remix, transform, and build upon the material for any purpose, even commercially.
- This license is acceptable for Free Cultural Works.
- The licensor cannot revoke these freedoms as long as you follow the license terms.

Referred to LAION-400M [16], Conceptual 12M [4] and MS-COCO [13], we only present the lists of URLs to this data. We build the meta file as follow.

[image_url] [class_index]

Problematic Content. The problematic contents such as drugs, nudity, and other offensive content exist in the web data. As mentioned in Sec. 1, the annotators were asked to discard such images instead of conducting annotation.

Privacy. To mitigate privacy issues with public visual datasets, researchers have attempted to obfuscate private information before publishing the data [7, 24]. We plan to follow this line of work to blur faces, license plates in our new annotated data. In addition, if the original picture found at the URL present on the OmniBenchmark on the record states users' names, phone numbers, or any personal information, users can request a takedown of this image.

Bias. The images were crawled from Flickr thus inheriting all the biases of that website. The usage of user generated data might bring the risk of bias. We plan to tackle this problem by balancing various categories.

Table 1. Top-1 accuracy of 26 models on 21 realms. We present the complete Top-1 accuracy of 25 models on 21 realms. ImageNet-1K-seen-realms are marked in underline. Consumer. indicates consumer_goods, Locom. indicates locomotive.

Method	Decoration	Creation	Process	Mammal	Instrument.	Material	Aquatic.	Activity.	Device	Amphibian	Bird
RN50 [9]	50.8	50.3	44.0	43.4	42.1	41.2	40.9	40.4	38.9	36.7	36.2
DINO [3]	54.2	56.7	48.4	46.1	44.7	46.3	45.3	47.4	42.4	38.6	38.4
MoCov2 [5]	50.5	52.4	45.4	41.5	41.1	41.8	40.5	42.3	38.7	36.5	31.1
SwAV [2]	54.8	56.5	47.9	45.8	44.1	45.7	44.0	47.1	42.1	37.9	36.5
BarlowTwins [27]	52.0	54.1	46.1	44.2	41.5	42.8	42.9	44.7	39.9	37.7	36.5
SwAV-Places [11]	47.6	51.8	46.3	32.2	37.6	39.8	37.0	43.4	36.7	29.0	21.6
MAE [8]	44.5	47.5	41.5	37.4	33.3	35.5	36.3	39.0	31.4	30.6	28.6
BeiT B/16 [1]	44.2	41.9	36.7	34.4	33.1	34.9	35.5	36.0	30.3	31.0	32.3
EfficientNetB4 [20]	50.3	52.3	44.5	45.0	45.5	43.2	44.5	40.9	41.9	38.9	38.0
MLP-Mixer [21]	48.6	47.8	42.0	40.1	40.6	40.0	40.4	38.9	36.5	33.8	30.2
SwinT [14]	61.8	62.2	52.8	54.9	57.6	55.0	55.3	53.1	54.5	46.7	53.9
ResNet-101 [9]	51.8	53.9	45.4	48.2	48.9	45.7	46.7	44.1	45.3	40.5	40.3
InceptionV4 [19]	47.2	48.8	38.7	43.2	44.5	39.3	39.7	37.5	39.8	36.8	32.5
MEAL-V2 [18]	52.5	53.7	45.6	46.9	47.7	44.7	43.2	43.4	43.7	38.3	36.2
CutMix [25]	47.7	48.2	40.3	42.6	43.1	39.3	35.4	36.8	38.6	35.1	30.3
Manifold [22]	47.9	48.8	41.2	42.6	41.9	39.0	38.1	37.7	37.4	34.6	31.6
ReLabel [26]	47.5	47.1	38.1	41.7	42.4	39.1	37.0	35.5	37.6	34.2	31.0
IG-1B [23]	57.3	58.2	49.6	49.4	50.5	48.6	47.2	47.1	46.6	42.4	43.2
GV-D [17]	57.4	59.8	52.1	50.6	48.9	51.3	50.9	52.2	47.5	45.3	52.9
IN21K [10]	56.9	57.6	48.0	46.8	47.5	47.5	47.6	48.1	44.9	41.1	45.3
CLIP [15]	59.5	62.0	54.0	44.9	49.1	49.8	42.5	53.7	46.3	38.5	39.6
MoPro-V1 [12]	51.6	52.7	45.9	45.6	43.8	43.3	42.1	42.7	40.6	38.7	37.5
ViT B/16 [6]	59.3	59.9	51.9	51.5	54.2	51.7	55.2	51.0	50.8	47.6	50.7
Peco RN50	51.4	53.5	45.0	45.2	43.4	43.6	43.5	43.1	41.3	39.3	38.8
Reco RN50	52.5	53.9	45.8	45.3	43.9	44.4	44.3	43.1	41.6	40.1	38.7
Peco RN101	54.0	54.8	47.3	48.0	46.5	45.7	46.0	45.0	43.6	39.4	41.1
Reco RN101	54.3	55.5	46.7	48.2	47.4	46.4	45.4	45.6	44.7	41.5	40.5

Method	Consumer.	Military.	Region	Aircraft	Structure	Locomot.	Geologi.	Plant.	Car	Food	AVG↑
RN50 [9]	32.5	32.0	28.8	27.6	26.9	26.0	25.7	25.1	18.1	12.7	34.3
DINO [3]	37.5	38.6	34.9	33.7	32.4	30.3	32.1	31.6	22.7	15.3	38.9
MoCov2 [5]	34.4	33.1	30.6	27.4	28.7	28.2	29.6	25.4	18.5	12.8	34.8
SwAV [2]	37.6	37.2	34.3	31.7	31.6	31.3	32.2	30.3	21.9	14.6	38.3
BarlowTwins [27]	35.6	37.2	32.8	32.6	30.2	29.6	30.3	28.4	22.3	14.2	36.9
SwAV-Places [2]	32.8	33.7	34.9	26.7	32.2	27.9	31.5	23.7	17.9	10.8	33.1
MAE [8]	29.2	29.4	26.7	26.6	25.1	23.7	25.4	22.8	17.4	10.0	30.6
BeiT B/16 [1]	26.5	24.6	26.4	20.4	22.8	22.4	24.0	23.0	13.8	8.5	30.1
EfficientNetB4 [20]	35.9	31.6	30.6	26.1	28.7	27.6	29.2	25.6	17.7	14.2	35.8
MLP-Mixer [21]	31.2	28.1	29.2	23.0	26.0	24.6	27.4	23.2	15.2	10.3	32.2
SwinT [14]	46.9	41.6	40.7	36.5	38.3	33.0	35.4	46.8	24.5	22.8	46.4
ResNet-101 [9]	38.1	33.0	32.2	27.8	30.1	25.3	28.9	27.0	18.7	14.2	37.4
InceptionV4 [19]	32.8	27.3	27.0	22.0	25.8	24.4	23.5	20.8	16.0	11.9	32.3
MEAL-V2 [18]	37.0	32.5	31.2	27.9	29.9	27.2	28.1	25.9	19.1	14.0	36.6
CutMix [25]	32.1	23.6	27.2	18.2	26.3	19.7	24.3	20.5	12.6	11.8	31.1
Manifold [22]	31.2	25.7	26.8	19.4	25.7	21.2	24.1	21.2	15.0	11.7	31.6
ReLabel [26]	30.3	25.5	25.9	19.5	24.9	21.3	23.0	20.7	13.5	11.8	30.8
IG-1B [23]	40.5	37.2	34.9	31.9	32.8	30.2	31.8	30.1	21.2	17.1	40.4
GV-D [17]	41.5	47.1	39.8	46.8	35.9	36.6	36.6	42.0	37.9	20.2	45.4
IN21K [10]	40.1	38.9	35.4	32.3	32.8	30.8	31.9	37.7	21.1	17.2	40.4
CLIP [15]	42.7	38.6	42.6	34.2	38.2	29.9	36.4	32.6	28.9	20.5	42.1
MoPro-V1 [12]	34.9	32.8	31.8	28.7	29.3	27.3	28.1	27.7	19.9	13.5	36.1
ViT B/16 [6]	44.2	40.2	37.9	34.5	36.7	30.1	33.3	44.3	23.5	21.6	45.8
Peco RN50	35.1	34.9	31.0	30.6	28.9	28.0	28.1	27.4	20.6	13.8	36.4
Reco RN50	35.5	35.1	31.4	31.1	29.1	28.6	28.6	27.8	20.6	14.1	36.9
Peco RN101	37.8	37.4	32.8	33.3	30.9	28.7	29.3	28.7	22.1	15.2	38.5
Reco RN101	38.4	36.9	32.7	33.8	31.0	28.4	28.9	28.6	21.8	15.5	38.7

Table 2. Accuracy of 26 models on 21 realms relative to the baseline ResNet50. We present the complete relative accuracy of 25 models on 21 realms. ImageNet-1K-seen-realms are marked in underline. Consumer. indicates consumer_goods, Locom. indicates locomotive..

Method	Decoration	Creation	Process	Mammal	Instrument	Material	Aquatic	Activity	Device	Amphibian	Bird
DINO [3]	3.4	6.4	4.4	2.7	2.6	5.1	4.4	7.0	3.4	1.9	2.2
MoCov2 [5]	-0.3	2.1	1.4	-2.0	-1.0	0.6	-0.4	1.9	-0.2	-0.2	-5.1
SwAV [2]	4.0	6.2	3.9	2.4	2.0	4.5	3.1	6.7	3.2	1.2	0.3
BarlowTwins [27]	1.2	3.7	2.2	0.8	-0.6	1.6	2.0	4.3	0.9	1.0	0.4
SwAV-Places [11]	-3.2	1.5	2.4	-11.2	-4.5	-1.5	-3.9	3.0	-2.3	-7.7	-14.6
MAE [8]	-6.3	-2.9	-2.5	-6.0	-8.8	-5.7	-4.6	-1.4	-7.6	-6.1	-7.6
BeiT B/16 [1]	-6.6	-8.5	-7.3	-9.0	-9.0	-6.3	-5.4	-4.4	-8.7	-5.7	-3.9
EfficientNetB4 [20]	-0.5	2.0	0.5	1.6	3.5	2.0	3.6	0.5	2.9	2.2	1.9
MLP-Mixer [21]	-2.2	-2.6	-2.0	-3.3	-1.5	-1.2	-0.5	-1.5	-2.5	-2.9	-5.9
SwinT [14]	11.0	11.9	8.8	11.5	15.6	13.8	14.4	12.7	15.6	10.0	17.7
ResNet-101 [9]	1.0	3.6	1.5	4.8	6.8	4.5	5.8	3.7	6.4	3.8	4.1
InceptionV4 [19]	-3.6	-1.6	-5.2	-0.2	2.4	-1.9	-1.2	-3.0	0.9	0.1	-3.7
MEAL-V2 [18]	1.7	3.3	1.7	3.5	5.7	3.5	2.3	3.0	4.8	1.6	0.1
CutMix [25]	-3.1	-2.2	-3.7	-0.8	1.0	-1.9	-5.5	-3.6	-0.3	-1.6	-5.8
Manifold [22]	-2.9	-1.5	-2.7	-0.9	-0.1	-2.2	-2.8	-2.7	-1.6	-2.1	-4.6
ReLabel [26]	-3.3	-3.3	-5.8	-1.8	0.3	-2.2	-3.9	-5.0	-1.4	-2.5	-5.1
IG-1B [23]	6.5	7.8	5.6	6.0	8.5	7.4	6.3	6.7	7.7	5.7	7.0
GV-D [17]	6.7	9.5	8.2	7.2	6.9	10.1	10.0	11.8	8.5	8.6	16.8
IN21K [10]	6.1	7.3	4.0	3.4	5.4	6.3	6.7	7.7	6.0	4.4	9.1
CLIP [15]	8.7	11.7	10.1	1.5	7.0	8.6	1.6	13.3	7.4	1.8	3.4
MoPro-V1 [12]	0.8	2.3	1.9	2.1	1.7	2.1	1.2	2.3	1.7	2.0	1.3
ViT B/16 [6]	8.6	9.6	8.0	8.1	12.1	10.5	14.3	10.6	11.8	10.9	14.5
PeCo RN50	0.6	3.2	1.0	1.8	1.3	2.4	2.6	2.7	2.3	2.6	2.7
ReCo RN50	1.7	3.6	1.9	1.9	1.8	3.2	3.4	2.7	2.7	3.4	2.5
PeCo RN101	3.3	4.5	3.3	4.6	4.4	4.5	5.2	4.6	4.7	2.7	5.0
ReCo RN101	3.5	5.2	2.7	4.8	5.3	5.2	4.5	5.2	5.7	4.8	4.4

Method	Consumer.	Military.	Region	Aircraft	Structure	Locomot.	Geologi.	Plant.	Car	Food	AVG↑
DINO [3]	5.0	6.6	6.1	6.1	5.5	4.3	6.4	6.5	4.7	2.7	4.6
MoCov2 [5]	1.9	1.2	1.8	-0.2	1.8	2.2	3.9	0.3	0.4	0.1	0.5
SwAV [2]	5.1	5.2	5.5	4.1	4.7	5.3	6.5	5.2	3.8	1.9	4.0
BarlowTwins [27]	3.1	5.2	4.0	5.0	3.3	3.6	4.6	3.2	4.2	1.5	2.6
SwAV-Places [2]	0.3	1.7	6.2	-0.8	5.3	1.9	5.8	-1.5	-0.2	-1.9	-1.2
MAE [8]	-3.3	-2.6	-2.1	-1.0	-1.7	-2.3	-0.3	-2.4	-0.7	-2.7	-3.7
BeiT B/16 [1]	-6.0	-7.4	-2.4	-7.2	-4.1	-3.6	-1.7	-2.2	-4.3	-4.2	-5.6
EfficientNetB4 [20]	3.4	-0.4	1.8	-1.5	1.8	1.6	3.5	0.5	-0.4	1.5	1.5
MLP-Mixer [21]	-1.3	-3.8	0.4	-4.6	-0.9	-1.4	1.7	-2.0	-2.9	-2.4	-2.1
SwinT [14]	14.4	9.7	12.0	9.0	11.4	7.0	9.7	21.7	6.5	10.2	12.1
ResNet-101 [9]	5.6	1.0	3.4	0.3	3.3	-0.7	3.2	1.9	0.7	1.6	3.2
InceptionV4 [19]	0.2	-4.7	-1.8	-5.6	-1.0	-1.6	-2.2	-4.4	-2.1	-0.7	-1.9
MEAL-V2 [18]	4.5	0.5	2.4	0.3	3.1	1.2	2.4	0.8	1.1	1.3	2.3
CutMix [25]	-0.4	-8.3	-1.6	-9.3	-0.6	-6.3	-1.5	-4.7	-5.5	-0.9	-3.2
Manifold [22]	-1.3	-6.3	-2.0	-8.1	-1.1	-4.8	-1.6	-4.0	-3.1	-1.0	-2.7
ReLabel [26]	-2.2	-6.5	-2.8	-8.1	-1.9	-4.7	-2.7	-4.4	-4.6	-0.9	-3.5
IG-1B [23]	8.0	5.2	6.2	4.3	6.0	4.2	6.1	5.0	3.1	4.4	6.1
GV-D [17]	9.0	15.2	11.1	19.2	9.0	10.6	10.9	16.9	19.9	7.6	11.1
IN21K [10]	7.6	7.0	6.6	4.8	5.9	4.8	6.2	12.5	3.1	4.6	6.2
CLIP [15]	10.2	6.7	13.8	6.7	11.4	3.9	10.7	7.5	10.9	7.9	7.8
MoPro-V1 [12]	2.4	0.8	3.0	1.2	2.4	1.3	2.4	2.6	1.8	0.9	1.8
ViT B/16 [6]	11.6	8.2	9.2	6.9	9.8	4.1	7.6	19.2	5.4	8.9	10.0
PeCo RN50	2.6	2.9	2.2	3.0	2.1	2.0	2.4	2.2	2.5	1.1	2.1
ReCo RN50	3.0	3.1	2.6	3.5	2.3	2.6	2.9	2.6	2.6	1.4	2.6
PeCo RN101	5.3	5.5	4.0	5.8	4.0	2.7	3.6	3.6	4.0	2.5	4.2
ReCo RN101	5.9	4.9	4.0	6.3	4.1	2.4	3.2	3.5	3.8	2.9	4.4

References

1. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers (2021) 3, 4
2. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. arXiv preprint arXiv:2006.09882 (2020) 3, 4
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers (2021) 3, 4
4. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: CVPR (2021) 2
5. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020) 3, 4
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 3, 4
7. Frome, A., Cheung, G., Abdulkader, A., Zennaro, M., Wu, B., Bissacco, A., Adam, H., Neven, H., Vincent, L.: Large-scale privacy protection in google street view. In: ICCV. pp. 2373–2380. IEEE (2009) 2
8. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners (2021) 3, 4
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 3, 4
10. Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., Houlsby, N.: Big transfer (bit): General visual representation learning (2020) 3, 4
11. Kotar, K., Ilharco, G., Schmidt, L., Ehsani, K., Mottaghi, R.: Contrasting contrastive self-supervised representation learning models. arXiv preprint arXiv:2103.14005 (2021) 3, 4
12. Li, J., Xiong, C., Hoi, S.C.: Mopro: Webly supervised learning with momentum prototypes. ICLR (2021) 3, 4
13. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755. Springer (2014) 2
14. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows (2021) 3, 4
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021) 3, 4
16. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs (2021) 2
17. Shao, J., Chen, S., Li, Y., Wang, K., Yin, Z., He, Y., Teng, J., Sun, Q., Gao, M., Liu, J., et al.: Intern: A new learning paradigm towards general vision. arXiv preprint arXiv:2111.08687 (2021) 1, 3, 4
18. Shen, Z., Savvides, M.: Meal v2: Boosting vanilla resnet-50 to 80accuracy on imagenet without tricks (2021) 3, 4
19. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning (2016) 3, 4

20. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks (2020) 3, 4
21. Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lucic, M., Dosovitskiy, A.: Mlp-mixer: An all-mlp architecture for vision (2021) 3, 4
22. Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., Bengio, Y.: Manifold mixup: Better representations by interpolating hidden states. In: ICML (2019) 3, 4
23. Yalniz, I.Z., Jégou, H., Chen, K., Paluri, M., Mahajan, D.: Billion-scale semi-supervised learning for image classification. arXiv preprint arXiv:1905.00546 (2019) 3, 4
24. Yang, K., Yau, J., Fei-Fei, L., Deng, J., Russakovsky, O.: A study of face obfuscation in imagenet. arXiv preprint arXiv:2103.06191 (2021) 2
25. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features (2019) 3, 4
26. Yun, S., Oh, S.J., Heo, B., Han, D., Choe, J., Chun, S.: Re-labeling imagenet: from single to multi-labels, from global to localized labels. arXiv preprint arXiv:2101.05022 (2021) 3, 4
27. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction (2021) 3, 4