A Experiment setups

In this section, we provide additional details of our experiments.

Unsupervised object proposals. When pretraining on the uncurated dataset, acquiring ground truth object bounding boxes using human annotations can be expensive. However, automatically generating unsupervised region proposal is well studied. We use Selective Search as the unsupervised proposal generation method. Following ORL [9] we first generate the proposals using selective search. Then we filter the proposals with 96 pixels as the minimal scale, maximum IOU of 0.5 and aspect ratio between 1/3 to 3. For every image we generate maximum of 100 proposals and randomly select any image as the object image.

OpenImages dataset. We use the OpenImages subset as proposed in [7]. This is a subset created from the full OpenImages dataset. In this subset each image has alteast 2 classes present and each class has atleast 900 instances. So this subset is a balanced subset of OpenImages with an average of 12 object present in an image. OpenImages subset is a very good proxy for mimicing real world multi-object images and hence we used it as the main dataset in our paper. More details of this dataset can be found in [7].

INPMix dataset. We sample a subset of the Place-205 [10] dataset to test the discriminative capacity of the representations on the scene images. Specifically, we randomly sample $1,300$ training images from each of the 205 classes, and then combine them with the ImageNet-100 [8] to form a dataset of 305 classes in total. We provide the code to reproduce the dataset in the supplementary material.

Object and Scene image augmentations. We find that small objects are always detrimental to the performance. Therefore, when sampling the objects using bounding boxes, we drop those bounding boxes with size $\texttt{width} \times \texttt{height} \leq 56 \times 56$. Further, when sampling objects for the Euclidean branch, if the size of a bounding box $\texttt{width} \times \texttt{height} \leq 256 \times 256$, we slightly expand it to either 256×256 or the maximal size allowed by the original image size. We also applied a small jittering to the `width` and `height` to include different contexts around the objects. Next, we apply random cropping and resizing with the same scale $(0.2, 1.)$ with MoCo [4]. Instead, when sampling objects for the hyperbolic branch, we do not apply jittering and random cropping, but only filtering the small boxes and resizing to $\leq 224 \times 224$. To crop the scene images, we sample another 1 to 5 bounding boxes and merge with the selected object bounding box.

B Additional experimental results

B.1 Robustness under Corruption.

We calculate the mCE error as calculated in Hendrycks et al. [5] . We compare our HCL model trained on OpenImages and lineval on ImageNet dataset with the baseline model without using HCL loss. We see an improvement of 1.9m CE over the baseline model, demonstrating that our HCL model learns more robust representations as compared to the vanilla MoCo.

B.2 Fine-grained class classification

Method	Cars [6]	DTD [2]	Food [1]
HCL/$\mathcal{L}_{\text{hyp}}$	31.92	68.46	58.66
HCL	32.02	68.19	58.79

Table 1: Fine grained classification results.

In Table 1 we show results on more fine grained classification results. We can see that on fine grained classification our model does not exhibit much performance improvement. This could be due to the fact that these datasets have more or less very similar context, hence the hyperbolic objective does not help too much in this case.

B.3 More ImageNet Examples

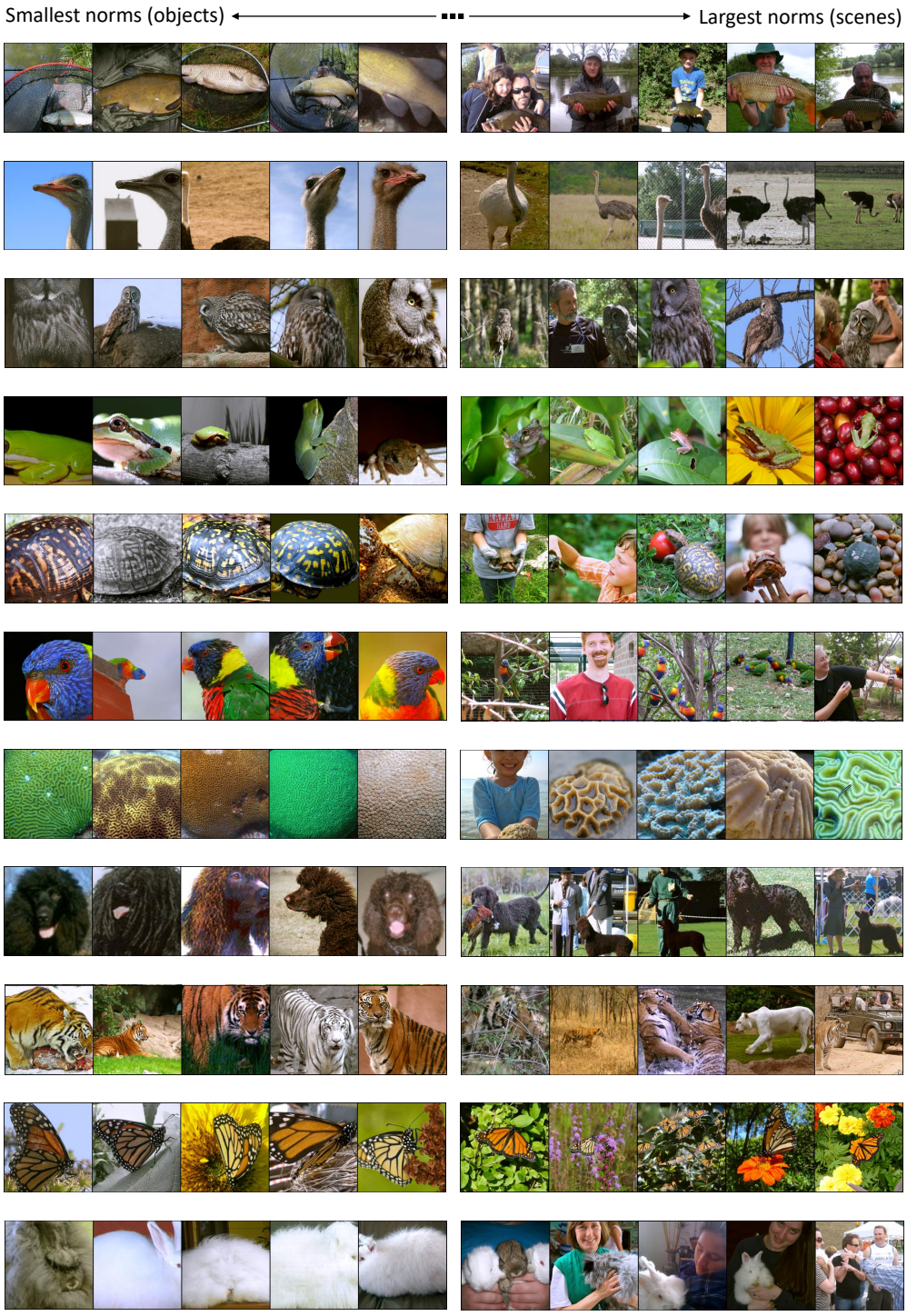

Figure 1: More images from ImageNet training set sorted by their representation norms.

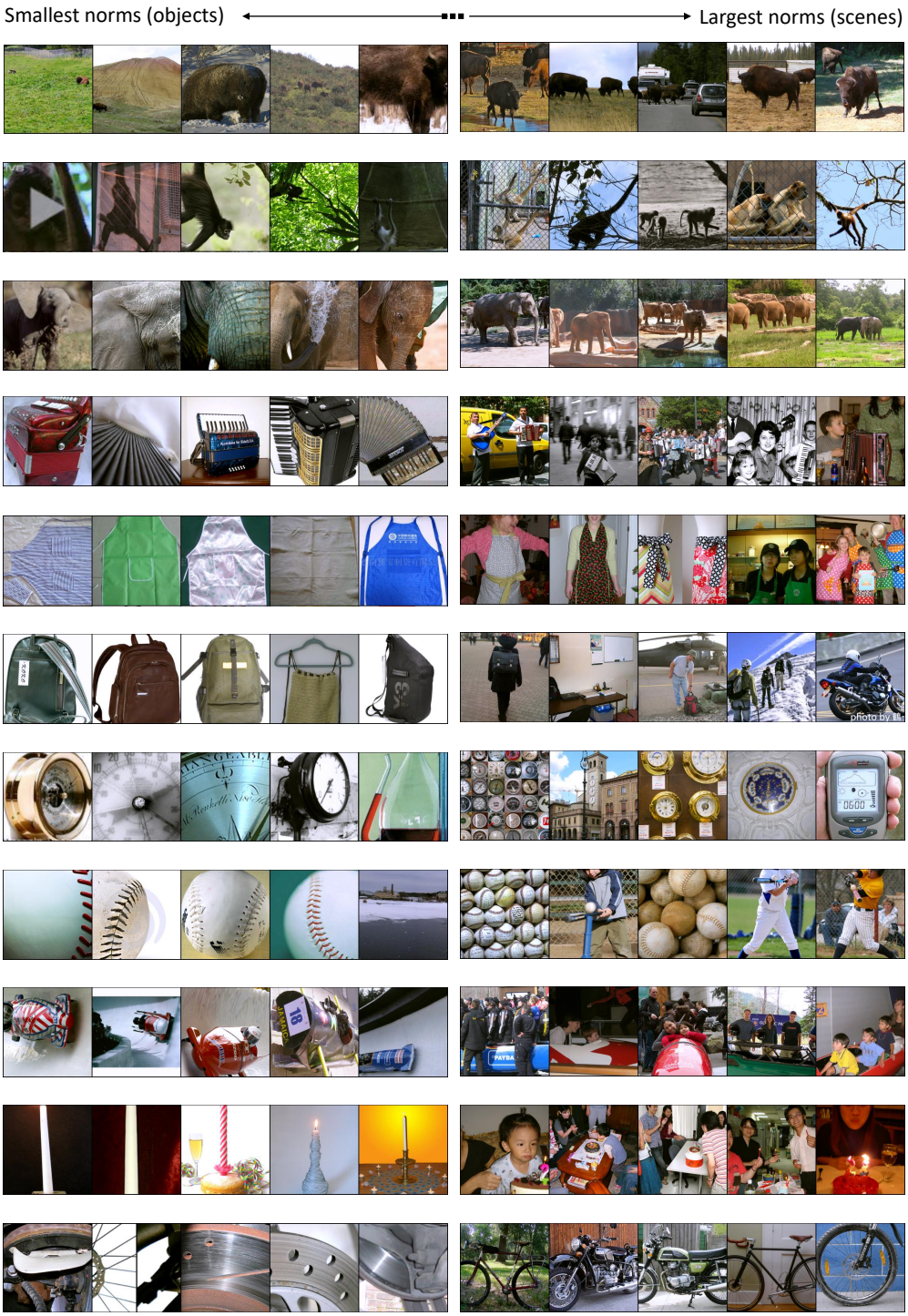

Figure 2: More images from ImageNet training set sorted by their representation norms.

C Additional ablation studies

In this section, we provide more ablation experiments on hyperbolic linear evaluation, model architecture, and the radius of Poincare ball. All the models are trained on the OpenImages dataset and evaluated on the ImageNet-100 (IN-100) or ImageNet-1k (IN-1k) the top-1 accuracy reported.

Radius of the Poincare ball In Table 2 we show results by varying the radius of Poincaré ball. The hyperbolic objective improves the performance over all the tested radius. We find that a too small radius may lead to a smaller improvement due to the stronger regularization.

Configuration of the encoder head In our experiments, the Euclidean and hyperbolic branches share the weights in both the backbone and the head of the encoders. We also try using a separate head for the hyperbolic branch. As shown in Table 3, this leads to a more stable training when larger learning rate is applied. However, we did not see any improvements brought by this modification.

Hyperbolic linear evaluation Apart from the common linear evaluation in the Euclidean space, we show the hyperbolic linear evaluation results with different optimizers and learning rates in Table 4. The idea is to test if the representations are more linearly separable in the hyperbolic space. We follow the same setting of hyperbolic softmax regression [3] and train a single hyperbolic linear layer. However, we find the optimization with SGD can easily causes to overflow. Instead, Adam works much more stable with appropriate learning rates.

c	IN-1k
1	58.08
0.5	58.31
0.1	58.29
0.05	58.51
0.01	58.49

Table 2: Results by varying the radius r of Poincaré ball. $c = \frac{1}{r^2}$.

Head	λ	IN-100
N/A	0	77.36
shared	0.1	79.08
	0.5	0
splitted	0.1	77.88
	0.5	77.58

Table 3: Different configurations of head in the the Euclidean and hyperbolic branches.

SGD		Adam	
lr	IN-100	lr	IN-100
0.1	63.82	0.001	67.64
0.2	64.22	0.0005	70.32
0.3	1	0.0001	72.58
0.4	1	0.00005	70.5

Table 4: Results of hyperbolic linear evaluation with different optimizers and learning rates.

References

[1] L. Bossard, M. Guillaumin, and L. Van Gool. Food-101 – mining discriminative components with random forests. In *ECCV*, 2014.

[2] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

[3] O. Ganea, G. Bécigneul, and T. Hofmann. Hyperbolic neural networks. *NeurIPS*, 2018.

[4] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[5] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ICLR*, 2019.

[6] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *3DRR*, 2013.

[7] S. K. Mishra, A. B. Shah, A. Bansal, A. N. Jagannatha, A. Sharma, D. Jacobs, and D. Krishnan. Object-aware cropping for self-supervised learning. *ArXiv*, abs/2112.00319, 2021.

[8] Y. Tian, D. Krishnan, and P. Isola. Contrastive multiview coding. In *ECCV*, 2020.

[9] J. Xie, X. Zhan, Z. Liu, Y. S. Ong, and C. C. Loy. Unsupervised object-level representation learning from scene images. In *NeurIPS*, 2021.

[10] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. *NeurIPS*, 2014.