Supplementary material for: Fine-grained Angular Contrastive Learning with Coarse Labels

Guy Bukchin[3,2], Eli Schwartz[1,2], Kate Saenko[1,4],
Ori Shahar[3], Rogerio Feris[1], Raja Giryes*[2], Leonid Karlinsky*[1]
IBM Research AI[1], Tel-Aviv University[2], Penta-AI[3], Boston University[4]

1. Additional baselines

We include 7 additional baseline comparisons using the four BREEDS [4] datasets. These baselines were not included in the main paper due to lack of space. The comparisons are summarized in Table 1. We include several kinds of additional baselines covering different possible supervised + self-supervised combinations, as well as of supervised contrastive loss use:

- **ensemble (of Coarse/Coarse+ with MoCoV2):** averaging the two models predictions (probabilities after softmax)
- **cascade (of Coarse/Coarse+ with MoCoV2):** classify into the max-scoring coarse classes $\mathcal{Y}_{coarse}$ using the coarse-supervised model (with the linear classifier resulting from pre-training), and then use the self-supervised model (with the LR classifier, Section 3.3 in the paper) for intra-class classification within the chosen coarse class (limiting the target classes to the sub-classes of the predicted coarse class)
- **concat (of Coarse/Coarse+ with MoCoV2):** concatenating the features produces by the two models, and doing few-shot classification via learning the logistic regression (paper Section 3.3) on the resulting concatenated features.
- **Supervised Contrastive:** Training with coarse labels $\mathcal{Y}_{coarse}$ using the Supervised Contrastive loss [3] replacing CE.

The "Coarse", "Coarse+", and "MoCoV2" models used in ensemble, cascade, and concat combinations are as described in the paper. As can be seen from the table, in all (the more challenging) all-way experiments ANCOR maintains a significant advantage over the baselines, even ones combining (via an ensemble, a cascade, or a concat) separately trained coarse-supervised and contrastive self-supervised models, and thus also using twice more learn-able parameters than ANCOR. This demonstrates once again the importance of joint coarse-supervised and contrastive self-supervised training employed by ANCOR and facilitated by the proposed angular normalization component enhancing the synergy between these two objectives.

*Equal contribution

	LIVING-17		NONLIVING-26		ENTITY-13		ENTITY-30	
Method	5-way	all-way	5-way	all-way	5-way	all-way	5-way	all-way
ensemble(Coarse+, MoCoV2)	74.07 ± 0.67	34.02 ± 0.13	73.14 ± 0.71	33.23 ± 0.11	85.27 ± 0.57	36.39 ± 0.08	84.22 ± 0.6	34.56 ± 0.08
ensemble(Coarse, MoCoV2)	84.32 ± 0.62	36.39 ± 0.13	79.95 ± 0.65	34.31 ± 0.11	85.97 ± 0.59	34.03 ± 0.08	88.48 ± 0.55	34.49 ± 0.08
cascade(Coarse+,MoCoV2)	74.1 ± 0.66	34.04 ± 0.13	73.1 ± 0.7	33.23 ± 0.11	85.19 ± 0.57	36.38 ± 0.08	84.33 ± 0.6	34.57 ± 0.08
cascade(Coarse,MoCoV2)	88.27 ± 0.63	38.02 ± 0.14	84.66 ± 0.64	36.05 ± 0.11	86.39 ± 0.59	35.46 ± 0.07	**90.02 ± 0.55**	37.1 ± 0.08
concat(Coarse+,MoCoV2)	73.48 ± 0.66	33.4 ± 0.13	71.49 ± 0.71	31.28 ± 0.11	84.26 ± 0.59	35.64 ± 0.08	83.03 ± 0.62	33.5 ± 0.07
concat(Coarse,MoCoV2)	83.45 ± 0.63	35.89 ± 0.13	77.59 ± 0.67	32.48 ± 0.11	85.35 ± 0.6	33.89 ± 0.07	86.83 ± 0.57	33.82 ± 0.07
Supervised Contrastive [3]	86.49 ± 0.67	35.11 ± 0.12	84.54 ± 0.61	37.44 ± 0.11	87.08 ± 0.58	28.57 ± 0.15	88.86 ± 0.52	33.67 ± 0.17
ANCOR (ours)	**89.23 ± 0.55**	**45.14 ± 0.12**	**86.23 ± 0.54**	**43.10 ± 0.11**	**90.58 ± 0.54**	**42.29 ± 0.08**	88.12 ± 0.54	**41.79 ± 0.08**

Table 1. **Additional baselines:** evaluated using BREEDS [4] with 1-shot. The "Coarse", "Coarse+", and "MoCoV2" models used in ensemble and cascade combinations are as described in the paper. (1) **ensemble:** averaging the two models predictions (probabilities after softmax); (2) **cascade:** classify into the max-scoring coarse classes $\mathcal{Y}_{coarse}$ using the coarse-supervised model (with the linear classifier resulting from pre-training), and then use the self-supervised model (with the LR classifier, paper Sec. 3.3) for intra-class classification within the chosen coarse class (limiting the target classes to its sub-classes); (3) **concat:** concatenate the feature vectors from the two models and then apply few-shot classification as in the paper Sec. 3.3; (4) Same as Coarse+ using the Supervised Contrastive loss [3] replacing CE.

2. Sub-population Shift

The 'sub-population shift' benchmark was proposed in [4] intended to evaluate how classification performance is affected when the train classes and test classes consist of different (non-overlapping set of) sub-classes (eg. 'Dog' class in training consist of samples of 'Bloodhound' and 'Pekinese', while the test dogs are the 'Great Pyrenese' and the 'papillon'). For this purpose they propose two hand-crafted partitions for each of their datasets: named 'good' and 'bad', which represent a less and a more adversarial partitioning respectively. We leverage this to create a task that is on one hand allows having the same coarse classes in training and in testing (as in the task in Section 4.4.1), while still having the sub-classes of those coarse classes non-overlapping (different) between the train and the test (as in the task in Section in 4.4.2). In other words, in this scenario the test sub-classes are completely different from the training ones, and yet they share the same parent coarse classes. Our evaluation on this task, provided in Table 2, shows that despite this challenging setting, ANCOR significantly outperforms the strongest baseline by $> 4\%$ in the all-way test.

	Good	Bad	5-way	all-way
Coarse+		✓	70.77 ± 0.74	36.65 ± 0.19
Coarse+	✓		73.44 ± 0.69	43.13 ± 0.22
ANCOR (ours)		✓	74.99 ± 0.71	40.69 ± 0.20
ANCOR (ours)	✓		**77.32 ± 0.69**	**47.53 ± 0.22**

Table 2. **Sub-population Shift**. Two hand-crafted partitions of the LIVING-17 dataset, created by [4], such that the test sub-classes are different from the (unlabeled) training sub-classes, yet share the common coarse classes. 'Good' and 'Bad' represent a less and more adversarial partitioning. Note that in practice these models train on half the data the models trained on LIVING-17 in the main paper have, due to the partitioning.

3. Additional results for 800-epoch training

In this section we further examine the effect of longer training longer on the performance. As can be seen in Table 3, following longer 800 epoch training ANCOR obtains significant gains in all the experimental settings. The most noticeable gains are in the all-way tests, where we observe that the gap above the baselines grows with longer training. We attribute this improvement to the contrastive component that is known to benefit from longer training [1, 2]. Interestingly, with longer training the coarse baseline models have gained accuracy in the 5-way test, but lost accuracy in the all-way test (compared to the 200 epochs performance). This supports our hypothesis that the coarse-classes supervised objective encourages reducing intra-class variation, and as such, with the longer training, tends to decresae the distinction between fine-sub classes loosing their discriminability within the coarse class. This again underlines the merits of our ANCOR approach that retains and enhances the fine sub-classes discrimnability thus significantly benefiting from the longer training regime.

	LIVING-17		NONLIVING-26		ENTITY-13		ENTITY-30	
Method	5-way	all-way	5-way	all-way	5-way	all-way	5-way	all-way
Fine (upper-bound)	91.94 ± 0.44	64.66 ± 0.17	87.68 ± 0.50	54.42 ± 0.13	94.22 ± 0.34	64.30 ± 0.09	93.11 ± 0.38	61.70 ± 0.09
Fine+ (upper-bound)	90.25 ± 0.48	63.54 ± 0.17	86.27 ± 0.51	53.16 ± 0.14	91.99 ± 0.40	59.43 ± 0.09	91.03 ± 0.43	57.48 ± 0.09
MoCoV2	81.45 ± 0.61	46.65 ± 0.16	78.33 ± 0.65	42.08 ± 0.12	87.30 ± 0.53	48.97 ± 0.08	86.51 ± 0.56	46.77 ± 0.09
MoCoV2-ImageNet [2]	89.27 ± 0.57	51.60 ± 0.15	82.22 ± 0.66	43.32 ± 0.12	88.30 ± 0.55	45.52 ± 0.08	87.18 ± 0.58	42.23 ± 0.08
SWAV-ImageNet [1]	80.11 ± 0.63	39.30 ± 0.14	73.43 ± 0.67	33.06 ± 0.11	79.58 ± 0.62	33.36 ± 0.07	78.89 ± 0.64	31.16 ± 0.07
Coarse	89.04 ± 0.63	29.06 ± 0.23	84.72 ± 0.63	27.99 ± 0.18	82.66 ± 0.75	11.24 ± 0.08	90.09 ± 0.59	20.63 ± 0.12
Coarse+	89.41 ± 0.61	33.07 ± 0.23	84.69 ± 0.59	32.07 ± 0.20	85.23 ± 0.63	22.85 ± 0.13	88.43 ± 0.55	28.33 ± 0.15
ANCOR (ours)	**92.59 ± 0.47**	**58.15 ± 0.16**	**88.25 ± 0.52**	**49.38 ± 0.13**	**92.04 ± 0.44**	**50.72 ± 0.09**	**92.13 ± 0.44**	**50.85 ± 0.09**

Table 3. Results for different baselines on the four BREEDS datasets. Every model was trained for 800 epochs.

4. Additional examples of ANCOR encoder $\mathcal{B}$ last layer activations

Additional examples of ANCOR encoder $\mathcal{B}$ last layer activations are provided in Figure 1, again illustrating an interesting attention to objects learned by ANCOR despite not being provided with any location supervision during training.

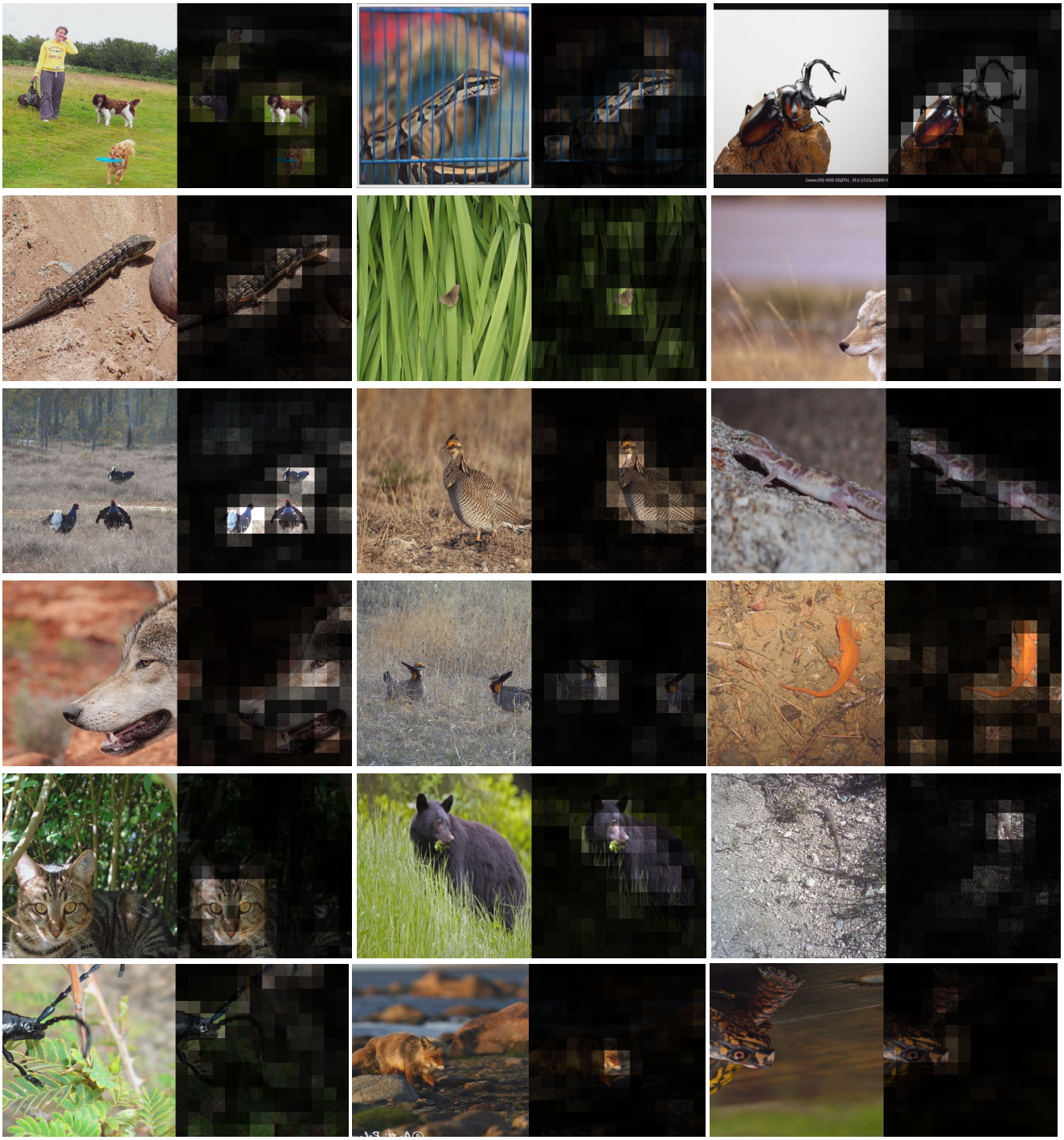

Figure 1. Additional examples of ANCOR encoder $\mathcal{B}$ last layer activations.

References

[1] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. 6 2020. 2

[2] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved Baselines with Momentum Contrastive Learning. *arXiv*, 3 2020. 2

[3] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised Contrastive Learning. *arXiv*, 4 2020. 1

[4] Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. BREEDS: Benchmarks for Subpopulation Shift. 8 2020. 1, 2