Supplementary material for *Concept Generalization in Visual Representation Learning*

Mert Bulent Sariyildiz[1,2,] Yannis Kalantidis[1] Diane Larlus[1] Karteek Alahari[2]
[1] NAVER LABS Europe [2] Inria[*]

Contents

This supplementary material is structured as follows. Sec. A details the design choices we made in creating the concept generalization levels of ImageNet-CoG (and is an extension of Sections 3.2 and 3.3 of the main paper). Sec. B describes the preprocessing pipeline for the models we benchmark in Sec. 4 of the main paper and also provides implementation details of our evaluation protocol (extending Sec. 3.4 of the main paper). Sec. C presents the complete set of results for the 31 models we evaluate in Sec. 4 of the main paper. Sec. D discusses how creating concept generalization levels with WordNet ontology and Lin similarity compares to creating them using the textual descriptions of concepts and a pretrained language model (and is an extension of Sec. 3.3 of the main paper). Finally, Sec. E discusses the impact of a recent update of ImageNet (impacting both ImageNet-21K and ImageNet-1K) on the results and future evaluation of our benchmark.

A. Details of ImageNet-CoG levels

We begin by detailing the steps to create the concept generalization levels of ImageNet-CoG. They include the selection of eligible unseen concepts in ImageNet-21K [11] (IN-21K) and the implementation details for Lin similarity [34]. We then briefly discuss the potential noise from missing labels in ImageNet-CoG. At the end of the section, we provide basic statistics about the ImageNet-CoG levels, i.e., the exact number of images per concept in each ImageNet-CoG level.

A.1. Selection criteria for unseen concepts

As described in Sec. 3.2 of the main paper, prior to creating the concept generalization levels, we determine a set of *eligible* unseen concepts in IN-21K. To determine these concepts, we implemented the following steps.

- We started with the whole set of IN-21K concepts (21,841) of the Fall 2011 release and excluded the ones from IN-1K, as they are the seen concepts.
- In order to create levels whose size is comparable to IN-1K, following the design choices made for IN-1K, we removed concepts with fewer than 782 images (note that any concept in IN-1K contains at least 782 images and 50 of those are used within the test set).
- It was shown that some of the concepts under the "person" sub-tree in IN-21K can be offensive or visually inappropriate, which may lead to undesirable behavior in downstream applications [66]. We thus excluded the entire "person" sub-tree.
- We also excluded all concepts that are parents in the ontology of eligible concepts, as this could lead to issues with the labeling strategy. Concretely, for any c_1 and c_2 in IN-21K, we exclude c_1 if c_1 is a parent of c_2.
- Finally, we manually inspected the remaining unseen concepts and found 70 potentially problematic concepts, which may be considered to be offensive, or too ambiguous to distinguish. Examples of such concepts include the very generic "People" (any group of human beings, men or women or children, collectively) or "Orphan" (a young animal without a mother) concepts. The list of such

[*] UGA, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France

manually discarded concepts is given in Tab. A.
After sequentially applying these steps, we are left with 5146 eligible unseen concepts. The complete list of eligible unseen concepts, along with the concepts in each level $L_{1/2/3/4/5}$ can be found on our project website.

A.2. Implementation details for Lin similarity

As described in Sec. 3.3 of the main paper, to produce the concept generalization levels, we sort the eligible unseen concepts by decreasing semantic similarity to the seen concepts in IN-1K. We used Lin similarity [34] as the semantic measure, which computes the relatedness of two concepts defined in a taxonomy.

Computing this similarity between two concepts requires their *information content* in the taxonomy. Following [44] and [46], we define the information content of a concept as $-\log p(c)$, where $p(c)$ is the probability of encountering concept c in the taxonomy. In our study, the taxonomy is a fragment of WordNet including all the concepts in IN-21K and their parents till the root node of WordNet: "Entity". Probability of a concept ranges between $[0, 1]$ such that if c_2 is a parent of c_1 then $p(c_1) < p(c_2)$, and the probability of "Entity" becomes 1.

In order to get superior-subordinate relationships between the concepts, we use WordNet-3.0 (the version ImageNet [11] is built on) implementation in the NLTK library [78].

A.3. Potential label noise in ImageNet-CoG

It has been shown recently [70] that ImageNet-1K (IN-1K) has missing-label noise. We can assume this extends to ImageNet-21K. Unfortunately, this type of noise is really difficult to correct and beyond the scope of our benchmark. However, we devise an experiment to get a sense of how much this noise could be. We take ResNet-50 classifiers for L_K and apply them to all the images of the IN-1K val set and vice versa (IN-1K classifiers on L_5 val). After inspecting samples that are predicted with very high confidence (> 0.99, about 2.7% of the images), we observe *several* cases where an unseen concept has (arguably) been seen during training without its label. Some examples are shown in Fig. A. Given the low percentage of very confident matches and the fact that [70] does not show a big change in performance after re-training with the noise corrected, we believe that this type of labeling noise does not significantly affect our findings.

A.4. Statistics for ImageNet-CoG

Number of images in each level. After selecting 1000 concepts for each level, we ensured that the image statistics are similar to those of IN-1K [47], i.e., we cap the number of images for each concept to a maximum of 1350 (1300 training + 50 testing). Note that we kept the same set of selected images per concept for all the experiments. We provide the complete list of image filenames in our code repository for reproducibility. In Fig. E, we plot the number of images per concept for each of the five levels and for IN-1K. We note a minor class imbalance in all the generalization levels from these plots. To investigate if this imbalance had any effect on the observations of our benchmark, we further evaluated a subset of the models analyzed in Sec. 4 of the main paper on a variant of the benchmark, where we randomly sub-sampled images from all the selected concepts to result in the same number of 732 training images, i.e., on class-balanced levels. Apart from the overall reduced accuracy as a result of smaller datasets, this experiment produced similar results to the ones shown in the main text, and all our observations continue to hold. We attribute this to the fact that imbalance is minimal.

B. Evaluation protocol of ImageNet-CoG

In this section, we provide additional implementation details of our evaluation protocol, thus extending Sec. 3.4 of the main paper.

B.1. Feature extraction and preprocessing

We establish evaluation protocols for ImageNet-CoG with image features extracted from pretrained visual backbones. To extract these features, we first resize an image such that its shortest side becomes S pixels, then take a center crop of size $S \times S$ pixels. To comply with the testing schemes of the models, for all the backbones we set $S = 224$, except *a*-Inception-v3 [53] ($S = 299$), *a*-DeiT-B-distilled [58] ($S = 384$), *a*-EfficientNet-B1 [54] ($S = 240$) and *a*-EfficientNet-B4 [54] ($S = 380$).

We also adapt their normalization schemes to be compatible with the data augmentation pipeline of the pretrained models. Concretely, we normalize each image by first dividing them by 255 (so that each pixel value is in $[0, 1]$), then applying mean and standard deviation normalization to the pixels, i.e., subtracting $[0.485, 0.456, 0.406]$ from the RGB channels and diving them by $[0.229, 0.224, 0.225]$, respectively. Note that for *d*-CLIP [43] we use mean $[0.481, 0.457, 0.408]$ and std $[0.268, 0.261, 0.275]$, and do not apply normalization for *s*-SimCLR-v2 [7].

Tab. B lists the set of unique backbone architectures considered in our study, and the dimensionality of the produced feature representations. For all the architectures trained in a supervised way, we extract features from the penultimate layers, i.e., before the last fully-connected layers making class predictions. For self-supervised learning methods, we follow the respective papers and extract features from the layer learned for transfer learning.

n00005787	n00288384	n00466377	n00466524	n00466630	n00474568	n00475014	n00475273	n00475403	n00483313
n00483409	n00483508	n00483848	n01314388	n01314663	n01314781	n01317294	n01317813	n01317916	n01318381
n01318894	n01321770	n01322221	n01323355	n01323493	n01323599	n01324431	n01324610	n01515303	n01517966
n01526521	n01862399	n01887474	n01888181	n02075612	n02152881	n02153109	n02156871	n02157206	n02236355
n02377063	n02377291	n02472987	n02475078	n02475669	n02759257	n02767665	n02771004	n03198500	n03300216
n03349771	n03393017	n03443005	n03680512	n04164406	n04193377	n04224543	n04425804	n04516354	n04979002
n06255081	n06272612	n06274760	n07942152	n08182379	n08242223	n08578517	n09828216	n10300303	n13918274

Table A: WordNet IDs of the 70 concepts considered problematic, therefore removed from the eligible list of unseen concepts.

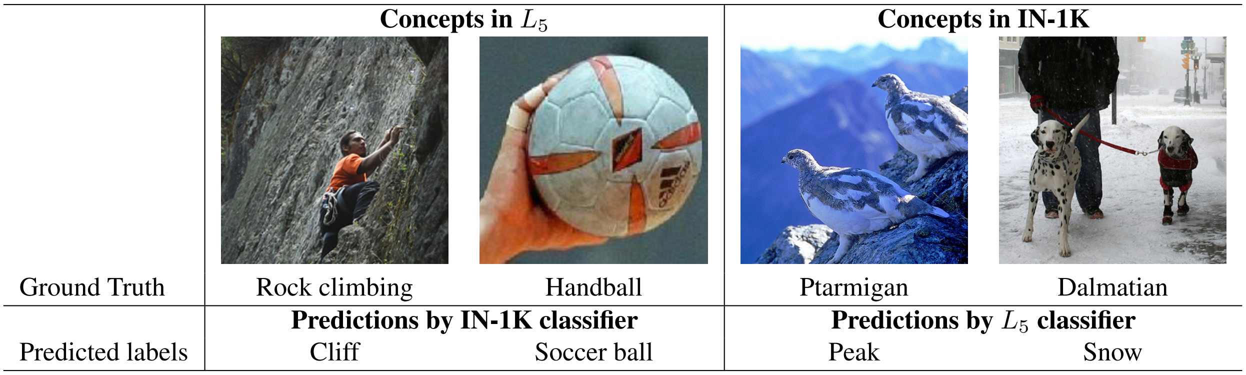


Figure A: **Illustration of the label noise in ImageNet-CoG.**

Model	Feature Dim.
All models with ResNet [23] backbone	2048
a-T2T-ViT-t-14 [68]	384
a-DeiT-S [58]	384
a-DeiT-B-distilled [58]	768
a-NAT-M4 [36]	1536
a-EfficientNet-B1 [54]	1280
a-EfficientNet-B4 [54]	1792
a-VGG19 [49]	4096

Table B: Unique architectures used by the models we evaluate in Sec. 4 of the main paper and the dimensionality of the feature vectors we extract from these architectures.

B.2. Training classifiers

In ImageNet-CoG, we perform two types of transfer learning experiments on each set of concepts, i.e., IN-1K or our concept generalization levels $L_{1/2/3/4/5}$ (see Sec. 4.2 of the main paper): (i) linear classification with all the available data, (ii) linear classification with a few randomly selected training samples. Both sets of experiments use the same test set, i.e., all the test samples.

In each of these experiments, we train a classifier with the features extracted using a given model. In order to evaluate each model in a fair manner in each setting, it is important to train each classifier in the best possible way.

We perform SGD to train classifiers, with momentum=0.9 updates, using batches of size 1024, and apply weight decay regularization to parameters. We choose learning rate and weight decay hyper-parameters on a validation set randomly sampled from the training set of each concept domain (20% of the training set is randomly sampled as a validation set for each concept domain). We sample 30 (learning rate, weight decay) pairs using Optuna [1] with a parzen estimator [77]. We then train the final classifier (with the hyper-parameters chosen from the previous validation step) on the full training set and report performance on the test set. We repeat this process 5 times with different seeds. This means that, in each repetition, we take a different random subset of the training set as a validation set and start hyper-parameter tuning with different random pairs of hyper-parameters. Despite this stochasticity, the overall pipeline is quite robust, with standard deviation in most cases less than 0.2, therefore, not clearly visible in figures. We will release training configurations along with our benchmark on our project website.

C. Extended results

C.1. Full set of results

In Sec. 4 of the main paper, we evaluate concept generalization performance for 31 models (listed in Tab. 1 of the main paper) on ImageNet-CoG. Figures 3 and 4 of the main paper report the results of training logistic regression classifiers with all the available training data for each concept (discussed in Sec. 4.2.1), and training it with a few samples per concept (discussed in Sec. 4.2.2), respectively. Due to space constraints, although Fig. 3 includes the results for all the models on all concept generalization levels, Fig. 4 provides only a selection of the few-shot results. In this section, we present the full set of results for all the methods when training with few and all data samples in table form. We also present the full set of figures for all the methods and levels when training with a few training samples per concept.

How fast can models adapt to unseen concepts. For completeness, we present the scores of all the models for $N = \{1, 2, 4, 8, 16, 32, 64, 128, \text{All}\}$ on IN-1K and $L_{1/2/3/4/5}$ in Fig. F (raw scores) and Fig. G (relative scores). These results, grouped by levels (i.e., for IN-1K and for $L_{1/2/3/4/5}$ separately) are also presented in Tables D, E, F, G, H, I respectively. These additional results complement Sec. 4.2.2 of the main paper.

Generalization to unseen concepts. To access the raw numbers of the results discussed in Sec. 4.2.1 of the main paper, we refer the reader to Tables D, E, F, G, H, I and the $N = \text{All}$ columns, which correspond to the scores shown in Fig. 3(a) of the main paper.

C.2. What if we fine-tuned the backbones?

Our benchmark and evaluation protocol are based on the assumption that good visual representations should generalize to different tasks with minimal effort. In fact, we explicitly chose to decouple representation learning and only consider frozen/pretrained backbones as feature extractors. We then evaluate how well pretrained representations transfer to concepts *unseen* during representation learning. Fine-tuning the models would therefore go against the main premise of our benchmark: after fine-tuning all concepts are "seen" during representation learning, i.e., the feature spaces can now be adapted. It would then be unclear: are we measuring the generalization capabilities of the pretraining strategy or of the fine-tuning process? How much does the latter affect generalization? We consider such questions out of the scope of our study. In fact, learning linear classifiers on top of pre-extracted features additionally allows us to exhaustively optimize hyper-parameters for all the methods and levels (see Section B.2), making sure that comparisons are fair across all models.

Measuring performance *relative to fine-tuning*, would however verify that the observed performance drops are due to increasing semantic distance and not variabilities across the levels. To this end, we fine-tune ResNet50 (pretrained on IN-1K) on IN-1K and on levels $L_{1/2/3/4/5}$ separately. Then we compare their performance with the protocol we chose for our benchmark, i.e. the case where we learn linear classifiers on top of pre-extracted features. In Fig. B, we show the *relative* scores of the linear classifiers on top of pre-extracted (labeled as "Pre-extracted") against fine-tuned ResNet50s (labeled as "Fine-tuned").

We observe that pre-extracted features become less and less informative for unseen concepts as we move from IN-1K to L_5, supporting our main assumption that semantically less similar concepts are harder to classify.

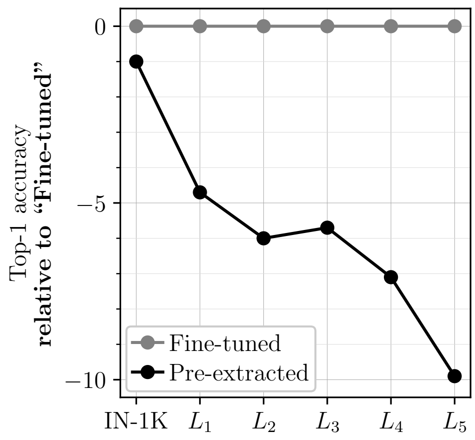


Figure B: Comparison of training linear classifiers on **pre-extracted features** *vs*. **fine-tuning** backbones on each level. Y-axis shows the top-1 accuracies obtained **relative** to the accuracy of the fine-tuned models.

D. An alternative semantic similarity

D.1. ImageNet-CoG with word2vecD

One of the requirements for studying concept generalization in a controlled manner is a knowledge base that provides the semantic relatedness of any two concepts. As IN-21K is built on the concept ontology of WordNet [41], in Sec. 3.3 of the main paper we leverage its graph structure, and propose a benchmark where semantic relationships are computed with the Lin measure [34].

As mentioned in Sec. 3 of the main paper, the WordNet ontology is hand-crafted, requiring expert knowledge. Therefore similarity measures that exploit this ontology (such as Lin) are arguably reliable in capturing the semantic similarity of concepts. However, it could also be desirable to learn semantic similarities automatically, for instance, using other knowledge bases available online such as Wikipedia. In this section, we investigate if such knowledge bases could be used in building our ImageNet-CoG.

With this motivation, we turn our attention to semantic similarity measures that can be learned over textual data describing the IN-21K concepts. Note that each IN-21K concept is provided with a name[1] and a short description.[2] The idea is to use this information to determine the semantic relatedness of any two concepts.

To this end, we leverage language models to map the

[1] http://www.image-net.org/archive/words.txt

[2] http://www.image-net.org/archive/gloss.txt

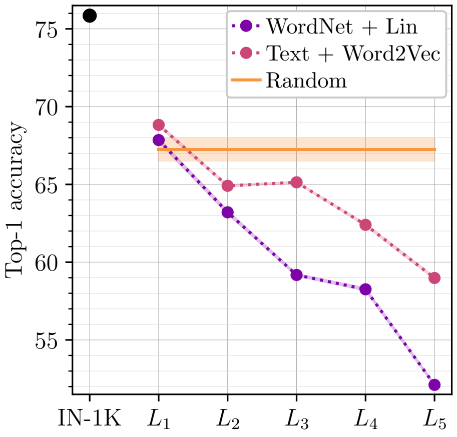


Figure C: **Semantic similarities** of the concepts captured by (i) Lin similarity [34] on WordNet graph [41] and (ii) cosine similarity of word2vec embeddings [79] extracted from textual descriptions of concepts, *vs.* **visual similarities** encoded by ResNet50, on IN-1K and generalization levels $L_{1/2/3/4/5}$ of ImageNet-CoG. We report the performance of linear logistic regression classifiers trained on features extracted from the global average pooling layer of ResNet50. The orange line shows results obtained on 1000 *random* unseen concepts (line represents the mean accuracy obtained over 15 random splits).

textual description of any concept into an embedding vector, such that the semantic similarity between two concepts can be measured as the similarity between their representations in that embedding space. We achieve this through the skip-gram language model [40], which has been extensively used in many natural language processing tasks, to extract "word2vec" representations of all concepts. However, we note that the name of many IN-21K concepts are *named entities* composed of multiple words, yet the vanilla skip-gram model tokenizes a textual sequence into words. We address this issue following [80] that learns a skip-gram model by taking into account such named entities. Specifically, we use the skip-gram model trained on Wikipedia[3] by the Wikipedia2Vec software [79].

We compute the word2vec embeddings of IN-21K concepts as follows. Firstly, we combine the names and descriptions of all concepts and learn tf-idf weights for each unique word. Secondly, for each concept, we compute two word2vec representations: one for the concept name, and one for the concept description, by averaging the word2vec representations of the words that compose them. These two average vectors are added and used as the final word2vec representation of the concept. Finally, as the semantic similarity measure, we simply use the cosine similarity between the word2vec representations of two concepts:

Dataset	Split	# Images
IN-1K [47]	train	1281167
IN-1K [47]	val	50000
IN-1K-blurred [81]	train	1281066
IN-1K-blurred [81]	val	49997

Table C: **Comparison of the number of images in IN-1K and IN-1K-blurred.**

$$\mathrm{sim}_{\mathrm{w2v}}(c_1, c_2) = \frac{\omega_{c_1}{}^\top \omega_{c_2}}{\|\omega_{c_1}\| \cdot \|\omega_{c_2}\|}, \tag{1}$$

where ω_c denotes the word2vec representation of concept c.

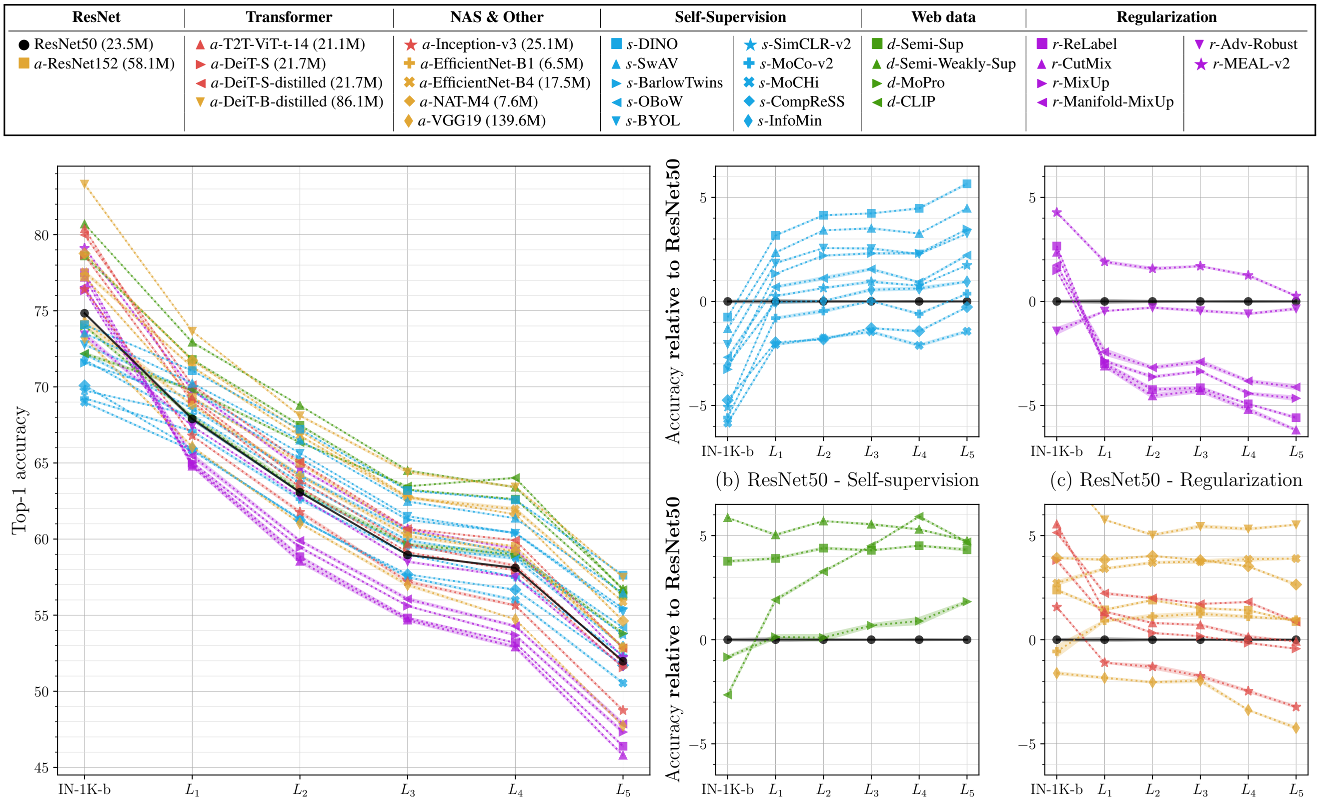
Recall that in Sec. 3.3 of the main paper, first we rank the 5146 eligible unseen concepts in IN-21K (which remain after our filtering, as explained in Sec. 3.3 of the main paper and Sec. A.1), w.r.t. their Lin similarity to the concepts in IN-1K. Then, we sub-sample 5000 concepts to construct concept generalization levels. To create another benchmark based on the textual information of the concepts as described above, we could repeat this procedure by replacing Lin similarity with the cosine similarity we defined in Eq. (1). However, this could select a different sub-set of 5000 concepts, which, in turn, would produce two benchmarks with different sets of unseen concepts. To prevent this, we re-rank the 5000 concepts selected by the Lin similarity, based on their text-based cosine similarity to IN-1K concepts. Then we simply divide the re-ordered concepts into 5 disjoint sequential sets.[4]

We compare the two benchmarks constructed with different knowledge bases (i.e., using the WordNet graph *vs.* textual descriptions) for our baseline model ResNet50 [23] that is pretrained on the seen concepts (IN-1K) for image classification, following our standard protocol. Concretely, first, we extract image features from the penultimate layer of the ResNet50, then we train linear classifiers on each concept domain separately.

We report results in Fig. C for the two benchmarks as well as randomly selected subsets of 1000 concept each. We see that the benchmark constructed using the WordNet ontology [41] and the Lin similarity [34] yield much more challenging concept generalization levels than the one obtained using textual data and a skip-gram language model [79] pretrained on Wikipedia. This is especially visible when comparing classification performance on the levels $L_{3/4/5}$ produced by each technique. We argue that this could be due to the fact that WordNet is an ontology hand-crafted by experts and is able to better approximate the semantic similarity of two concepts compared to the learned skip-gram model.

[3] April 2018 version of the English Wikipedia dump.

[4] Note that, given that the percentage of discarded concepts is very small (less than 3%, as 146 concepts are discarded from the 5146 eligible ones), this choice has minimal impact anyway.



(a) Top-1 accuracy across concept generalization levels

(b) ResNet50 - Self-supervision

(c) ResNet50 - Regularization

(d) ResNet50 - Web data

(e) Backbone architecture

Figure D: **Linear classification on ImageNet-CoG using blurred images for IN-1K.** Top-1 accuracies for all the 31 models listed in Tab. 1 of the main paper, after training logistic regression classifiers on the **blurred version** of IN-1K (IN-1K-b in the plots) and each level $L_{1/2/3/4/5}$. (a) Absolute top-1 accuracy on all levels. (b)-(e) accuracy relative to the baseline ResNet50 for all the models, split across the four model categories presented in Sec. 4.1 of the main paper.

We see that, for a given level L_i, WordNet combined with Lin similarity manages to gather concepts that are harder to discriminate and that the resulting classification performance is lower. This experiment, however, shows that it is possible to create a similar benchmark using automatically produced semantic similarity scores, the main alternative in the absence of any reliable hand-crafted ontology.

E. The ImageNet 2021 release

The ImageNet team recently released a new version of the full ImageNet dataset (IN-21K) as well as the ILSVRC-2012 dataset (IN-1K).[5] with this release, both the datasets are now available for download directly from the official website.[6]

The 2021 Winter version of IN-21K. We built ImageNet-CoG on the 2011 Fall release of IN-21K, which was the only version available in 2020, when we started constructing our benchmark. The 2011 Fall version contained 21841 concepts, while the new release has only 19167 concepts–a subset of the concepts from the Fall 2011 release. This follows recent studies from the ImageNet team, which identify potentially problematic concepts [66]. They were removed from the latest ImageNet version, including all the concepts under the “Person” sub-tree in WordNet.

With this modified version we successfully verified that: i) all the concepts of ImageNet-CoG are available in the new release, and ii) the images for all the 5000 concepts of ImageNet-CoG are identical in both releases. Consequently, all the results in our work *can also be reproduced* using the Winter 2021 version of IN-21K.

Blurred version of IN-1K. To protect the privacy of people present in some of the IN-1K images, the ImageNet team released a new version of this dataset, which we refer to as IN-1K-blurred [81]. In this version, the faces of people are blurred in the images. The statistics of these two versions are compared in Tab. C.

Although the models we evaluated in the paper were pretrained on IN-1K, with non-blurred images, for future reference, we performed our evaluation also on the blurred

[5] https://image-net.org/update-mar-11-2021.php

[6] https://image-net.org/download-images.php

version of IN-1K (IN-1K-blurred) for all the models. Concretely, for each model, we follow our evaluation protocol on IN-1K-blurred by extracting features of the blurred images and training logistic regression classifiers on them. We report these results in Fig. D. Note that Fig. D is the new version of Fig.3 in the main paper, with results obtained on IN-1K-blurred instead of IN-1K. We observe that the scores drop on average 0.91%, which is comparable to the 0.68% drop observed on popular models [81].

References

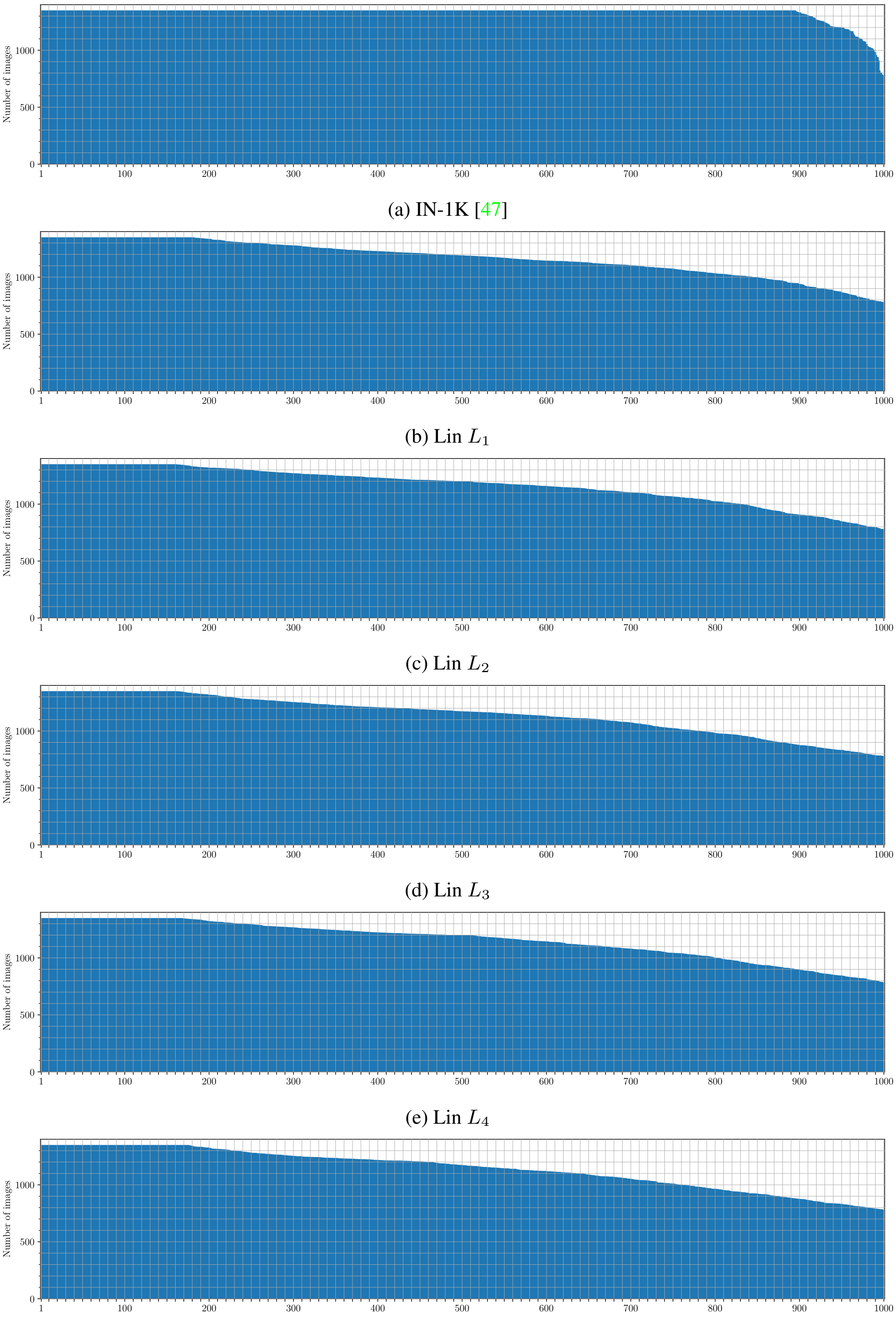
[77] James S Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Proc. NeurIPS*, 2011. 3

[78] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. 2009. 2

[79] Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. Wikipedia2vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from wikipedia. In *Proc. EMNLP*, 2020. 5

[80] Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. Joint learning of the embedding of words and entities for named entity disambiguation. In *Proc. CONLL*, 2016. 5

[81] Kaiyu Yang, Jacqueline Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. A study of face obfuscation in ImageNet. *arXiv preprint arXiv:2103.06191*, 2021. 5, 6, 7



(a) IN-1K [47]

(b) Lin L_1

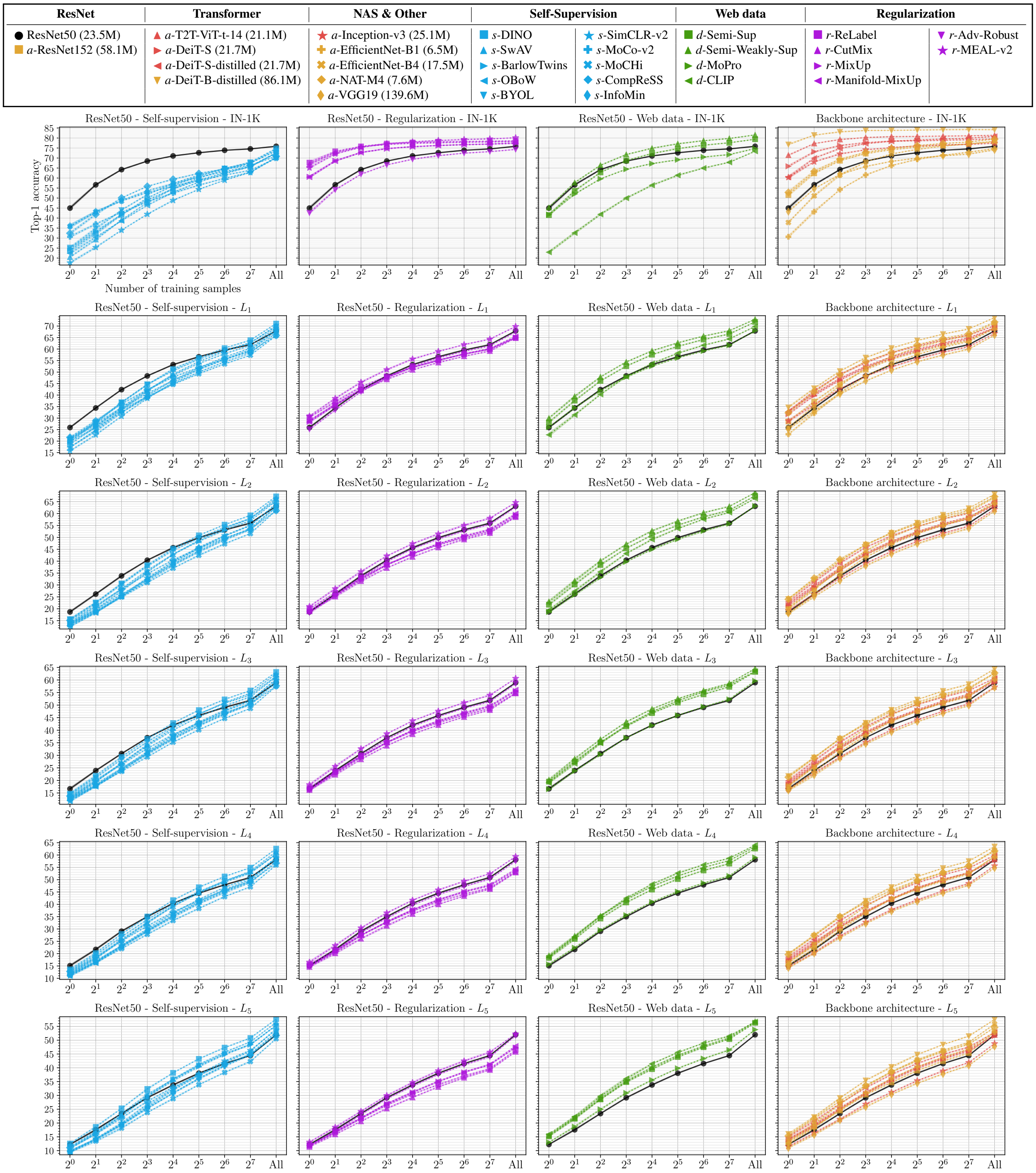
(c) Lin L_2

(d) Lin L_3

(e) Lin L_4

(f) Lin L_5

Figure E: The number of images per concept for IN-1K [47] and each of the concept generalization levels obtained by Lin similarity. We end up with 1.17M, 1.17M, 1.15M, 1.16M, 1.14M images in total for levels $L_{1/2/3/4/5}$ respectively. Note that IN-1K has 1.33M images in total.



ResNet50 - Self-supervision - IN-1K
ResNet50 - Regularization - IN-1K
ResNet50 - Web data - IN-1K
Backbone architecture - IN-1K
ResNet50 - Self-supervision - L_1
ResNet50 - Regularization - L_1
ResNet50 - Web data - L_1
Backbone architecture - L_1
ResNet50 - Self-supervision - L_2
ResNet50 - Regularization - L_2
ResNet50 - Web data - L_2
Backbone architecture - L_2
ResNet50 - Self-supervision - L_3
ResNet50 - Regularization - L_3
ResNet50 - Web data - L_3
Backbone architecture - L_3
ResNet50 - Self-supervision - L_4
ResNet50 - Regularization - L_4
ResNet50 - Web data - L_4
Backbone architecture - L_4
ResNet50 - Self-supervision - L_5
ResNet50 - Regularization - L_5
ResNet50 - Web data - L_5
Backbone architecture - L_5

Figure F: **Few-shot linear classification on ImageNet-CoG.** Top-1 accuracy for each method using logistic regression classifiers. We train them on pre-extracted features for the concepts in IN-1K and our generalization levels ($L_{1/2/3/4/5}$), with a few training samples per concept, i.e., $N = \{1, 2, 4, 8, 16, 32, 64, 128\}$. "All", the performance when all the samples are used, is also shown for reference.

ResNet	Transformer	NAS & Other	Self-Supervision		Web data	Regularization	
● ResNet50 (23.5M)	▲ *a*-T2T-ViT-t-14 (21.1M)	★ *a*-Inception-v3 (25.1M)	■ *s*-DINO	★ *s*-SimCLR-v2	■ *d*-Semi-Sup	■ *r*-ReLabel	▼ *r*-Adv-Robust
■ *a*-ResNet152 (58.1M)	▶ *a*-DeiT-S (21.7M)	✚ *a*-EfficientNet-B1 (6.5M)	▲ *s*-SwAV	✚ *s*-MoCo-v2	▲ *d*-Semi-Weakly-Sup	▲ *r*-CutMix	★ *r*-MEAL-v2
	◀ *a*-DeiT-S-distilled (21.7M)	✖ *a*-EfficientNet-B4 (17.5M)	▶ *s*-BarlowTwins	✖ *s*-MoCHi	▶ *d*-MoPro	▶ *r*-MixUp	
	▼ *a*-DeiT-B-distilled (86.1M)	◆ *a*-NAT-M4 (7.6M)	◀ *s*-OBoW	◆ *s*-CompReSS	◀ *d*-CLIP	◀ *r*-Manifold-MixUp	
		♦ *a*-VGG19 (139.6M)	▼ *s*-BYOL	♦ *s*-InfoMin			

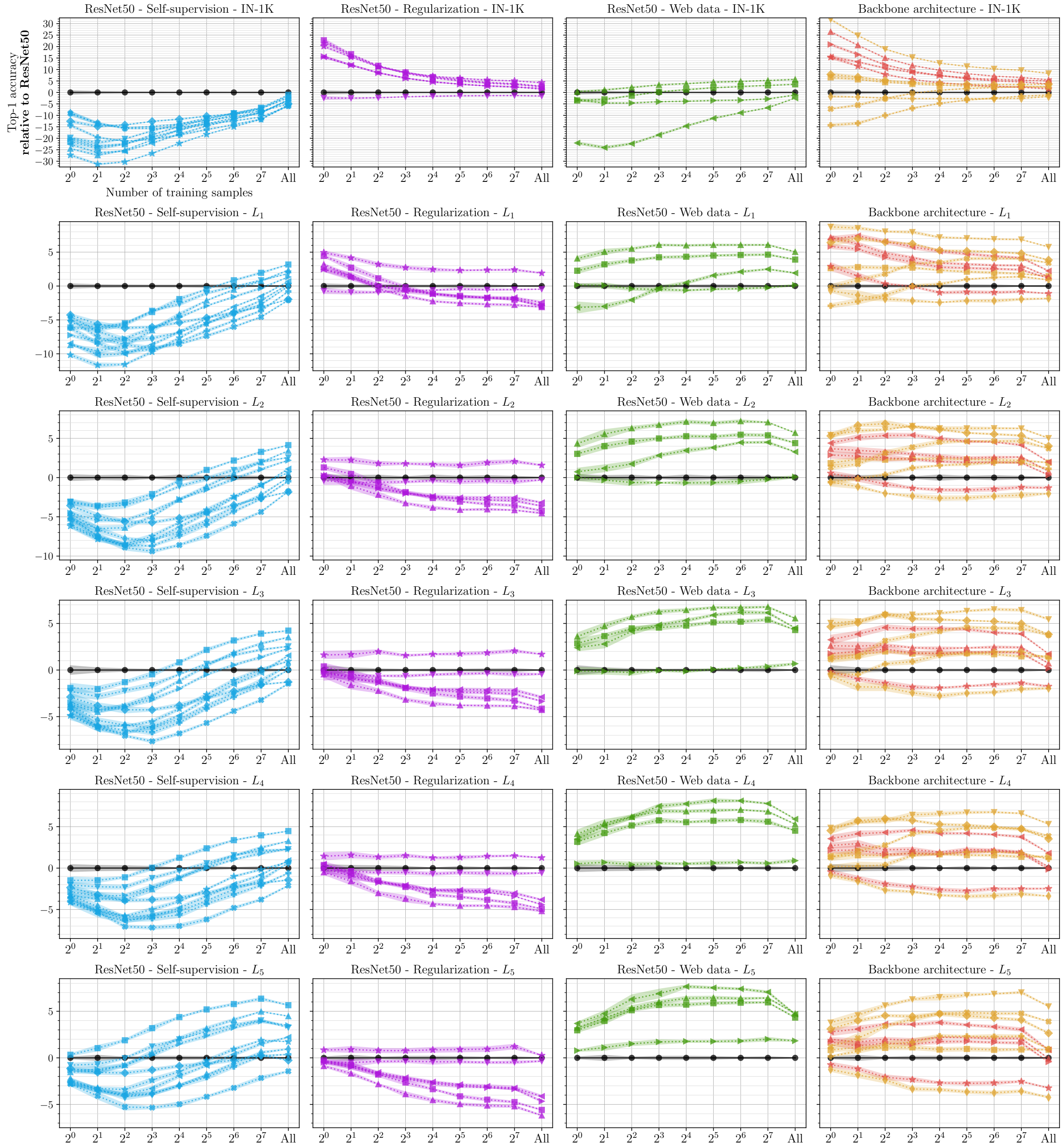


Figure G: **Relative few-shot linear classification on ImageNet-CoG.** The scores shown in Fig. F from a different perspective: all scores are **relative to ResNet50**.

	N-shots								
Model	2^0	2^1	2^2	2^3	2^4	2^5	2^6	2^7	All
ResNet50	45.0 +- 0.7	56.6 +- 0.4	64.2 +- 0.2	68.5 +- 0.1	71.0 +- 0.0	72.6 +- 0.1	73.9 +- 0.1	74.6 +- 0.1	75.8 +- 0.0
a-ResNet-152	51.5 +- 0.7	62.3 +- 0.3	68.7 +- 0.1	72.2 +- 0.1	74.2 +- 0.1	75.4 +- 0.1	76.3 +- 0.1	77.0 +- 0.1	78.1 +- 0.1
a-T2T-ViTt-14	71.4 +- 0.3	77.2 +- 0.1	79.2 +- 0.1	80.2 +- 0.1	80.7 +- 0.1	80.8 +- 0.0	80.9 +- 0.0	81.1 +- 0.1	81.3 +- 0.0
a-DeiT-S	65.9 +- 0.5	73.1 +- 0.3	76.2 +- 0.2	77.7 +- 0.0	78.4 +- 0.1	78.6 +- 0.1	78.9 +- 0.0	79.1 +- 0.1	79.6 +- 0.0
a-DeiT-S distilled	60.4 +- 0.7	70.0 +- 0.2	74.8 +- 0.2	77.2 +- 0.1	78.4 +- 0.2	79.0 +- 0.1	79.5 +- 0.0	79.9 +- 0.1	80.8 +- 0.0
a-DeiT-B distilled	76.8 +- 0.3	81.4 +- 0.1	83.1 +- 0.2	83.8 +- 0.1	83.8 +- 0.0	84.0 +- 0.0	84.1 +- 0.1	84.2 +- 0.0	84.2 +- 0.0
a-Inception-v3	60.3 +- 0.7	68.1 +- 0.2	72.0 +- 0.2	74.1 +- 0.1	75.2 +- 0.1	76.0 +- 0.1	76.5 +- 0.1	76.8 +- 0.1	77.4 +- 0.0
a-EfficientNet-B1	30.7 +- 0.6	43.2 +- 0.6	54.2 +- 0.1	61.6 +- 0.2	66.2 +- 0.1	69.3 +- 0.2	71.4 +- 0.1	72.9 +- 0.1	74.7 +- 0.3
a-EfficientNet-B4	37.8 +- 0.3	51.1 +- 0.4	61.5 +- 0.5	68.1 +- 0.1	72.0 +- 0.1	74.3 +- 0.1	76.0 +- 0.1	77.1 +- 0.1	78.4 +- 0.1
a-NAT-M4	52.8 +- 0.7	63.3 +- 0.5	69.4 +- 0.2	72.8 +- 0.2	74.8 +- 0.1	76.2 +- 0.1	77.3 +- 0.1	78.0 +- 0.0	79.5 +- 0.1
a-VGG19	43.2 +- 0.4	54.6 +- 0.1	61.8 +- 0.2	65.8 +- 0.1	68.3 +- 0.1	69.9 +- 0.1	71.1 +- 0.1	72.0 +- 0.1	74.1 +- 0.1
s-DINO	23.6 +- 0.5	32.2 +- 0.7	41.6 +- 0.2	49.9 +- 0.2	56.2 +- 0.3	60.9 +- 0.1	64.7 +- 0.1	67.9 +- 0.2	74.8 +- 0.0
s-SwAV	20.5 +- 0.3	29.3 +- 0.4	39.0 +- 0.1	47.7 +- 0.1	54.8 +- 0.3	60.0 +- 0.1	64.1 +- 0.1	67.5 +- 0.1	74.3 +- 0.0
s-BarlowTwins	24.7 +- 0.6	33.3 +- 0.6	41.7 +- 0.2	49.0 +- 0.2	54.6 +- 0.2	59.0 +- 0.1	62.7 +- 0.1	65.7 +- 0.1	72.3 +- 0.0
s-OBoW	22.4 +- 0.6	30.4 +- 0.3	38.7 +- 0.2	46.4 +- 0.2	52.8 +- 0.2	58.0 +- 0.1	62.2 +- 0.1	65.7 +- 0.1	72.7 +- 0.0
s-BYOL	25.2 +- 0.5	34.7 +- 0.6	44.0 +- 0.2	51.7 +- 0.2	57.2 +- 0.2	61.4 +- 0.1	64.8 +- 0.2	67.5 +- 0.1	73.5 +- 0.0
s-SimCLR-v2	17.7 +- 0.5	25.3 +- 0.3	33.9 +- 0.1	41.9 +- 0.2	48.8 +- 0.2	54.3 +- 0.1	58.9 +- 0.1	62.8 +- 0.0	70.5 +- 0.0
s-MoCo-v2	30.6 +- 0.6	37.0 +- 0.3	43.0 +- 0.1	48.0 +- 0.2	52.5 +- 0.3	56.4 +- 0.2	59.8 +- 0.1	62.9 +- 0.2	70.1 +- 0.1
s-MoCHi	35.8 +- 0.9	43.1 +- 0.5	48.5 +- 0.2	52.7 +- 0.1	55.9 +- 0.3	58.9 +- 0.2	61.4 +- 0.1	63.9 +- 0.1	69.9 +- 0.1
s-CompReSS	32.4 +- 0.6	41.8 +- 0.5	50.1 +- 0.1	55.8 +- 0.1	59.4 +- 0.2	62.3 +- 0.2	64.5 +- 0.1	66.4 +- 0.1	70.9 +- 0.0
s-InfoMin	35.9 +- 0.8	43.1 +- 0.4	48.8 +- 0.1	53.6 +- 0.2	57.2 +- 0.3	60.4 +- 0.1	63.3 +- 0.1	65.9 +- 0.1	72.5 +- 0.0
d-Semi-Sup.	41.5 +- 0.6	53.6 +- 0.4	62.8 +- 0.1	68.7 +- 0.1	72.2 +- 0.2	74.7 +- 0.1	76.4 +- 0.1	77.6 +- 0.1	79.4 +- 0.0
d-Semi-Weakly-Sup.	45.2 +- 0.5	57.7 +- 0.2	66.4 +- 0.2	71.7 +- 0.1	74.9 +- 0.1	77.1 +- 0.2	78.7 +- 0.1	79.8 +- 0.1	81.5 +- 0.0
d-MoPro	41.8 +- 0.5	52.0 +- 0.2	59.6 +- 0.1	64.4 +- 0.1	67.2 +- 0.1	69.1 +- 0.1	70.5 +- 0.1	71.7 +- 0.1	74.7 +- 0.0
d-CLIP	22.9 +- 0.5	32.6 +- 0.5	41.9 +- 0.4	49.9 +- 0.3	56.4 +- 0.3	61.4 +- 0.3	65.0 +- 0.1	67.9 +- 0.1	73.4 +- 0.0
r-ReLabel	67.7 +- 0.8	73.3 +- 0.2	75.8 +- 0.0	77.2 +- 0.1	77.8 +- 0.1	78.1 +- 0.1	78.2 +- 0.1	78.4 +- 0.0	78.6 +- 0.0
r-CutMix	66.8 +- 0.5	72.7 +- 0.2	75.6 +- 0.1	76.8 +- 0.1	77.4 +- 0.1	77.6 +- 0.1	77.9 +- 0.1	78.0 +- 0.1	78.3 +- 0.0
r-Mixup	60.6 +- 0.4	68.6 +- 0.3	72.8 +- 0.2	74.7 +- 0.1	75.6 +- 0.1	76.2 +- 0.0	76.7 +- 0.0	76.9 +- 0.0	77.3 +- 0.0
r-Manifold Mixup	60.5 +- 0.5	68.5 +- 0.1	72.7 +- 0.2	74.7 +- 0.1	75.7 +- 0.0	76.3 +- 0.1	76.7 +- 0.1	77.0 +- 0.1	77.7 +- 0.0
r-AdvRobust	42.6 +- 0.7	54.1 +- 0.2	61.9 +- 0.1	66.6 +- 0.1	69.4 +- 0.1	71.2 +- 0.1	72.4 +- 0.1	73.2 +- 0.1	74.3 +- 0.1
r-MEAL-v2	65.1 +- 0.5	72.1 +- 0.2	75.4 +- 0.1	77.2 +- 0.2	78.1 +- 0.1	78.8 +- 0.2	79.3 +- 0.1	79.6 +- 0.1	80.1 +- 0.0

Table D: **Top-1 accuracies obtained by linear classifiers on IN-1K.** Table view corresponding to the 1st row in Fig. F.

	N-shots								
Model	2^0	2^1	2^2	2^3	2^4	2^5	2^6	2^7	All
ResNet50	25.9 +- 0.3	34.4 +- 0.2	42.4 +- 0.2	48.3 +- 0.1	53.2 +- 0.2	56.7 +- 0.2	59.6 +- 0.2	61.9 +- 0.1	67.9 +- 0.1
a-ResNet-152	28.5 +- 0.2	37.2 +- 0.3	45.1 +- 0.2	51.0 +- 0.2	55.6 +- 0.1	58.8 +- 0.2	61.5 +- 0.2	63.7 +- 0.1	69.3 +- 0.0
a-T2T-ViTt-14	33.1 +- 0.3	40.6 +- 0.3	47.2 +- 0.3	52.5 +- 0.2	56.6 +- 0.1	59.9 +- 0.1	62.6 +- 0.1	64.9 +- 0.1	69.2 +- 0.0
a-DeiT-S	31.7 +- 0.3	39.8 +- 0.3	46.6 +- 0.2	51.9 +- 0.3	56.0 +- 0.2	59.3 +- 0.2	62.1 +- 0.1	64.3 +- 0.1	69.0 +- 0.0
a-DeiT-S distilled	32.8 +- 0.4	41.8 +- 0.3	48.9 +- 0.3	54.1 +- 0.3	58.4 +- 0.2	61.4 +- 0.2	64.0 +- 0.2	66.1 +- 0.1	70.1 +- 0.1
a-DeiT-B distilled	34.6 +- 0.4	42.9 +- 0.2	50.4 +- 0.1	56.3 +- 0.2	60.4 +- 0.1	63.7 +- 0.1	66.5 +- 0.1	68.8 +- 0.0	73.7 +- 0.0
a-Inception-v3	28.8 +- 0.6	35.9 +- 0.5	42.7 +- 0.2	48.3 +- 0.1	52.3 +- 0.2	55.8 +- 0.2	58.6 +- 0.2	61.1 +- 0.1	66.8 +- 0.0
a-EfficientNet-B1	23.0 +- 0.3	32.1 +- 0.4	41.0 +- 0.4	48.3 +- 0.2	53.5 +- 0.2	57.6 +- 0.1	60.8 +- 0.1	63.3 +- 0.1	68.8 +- 0.0
a-EfficientNet-B4	25.7 +- 0.4	35.3 +- 0.3	44.1 +- 0.2	51.6 +- 0.2	56.7 +- 0.1	60.7 +- 0.2	63.6 +- 0.2	65.9 +- 0.1	71.3 +- 0.0
a-NAT-M4	32.3 +- 0.5	41.4 +- 0.3	48.9 +- 0.1	54.6 +- 0.1	58.5 +- 0.1	61.8 +- 0.2	64.6 +- 0.1	66.7 +- 0.1	71.7 +- 0.0
a-VGG19	25.2 +- 0.2	33.0 +- 0.7	40.5 +- 0.3	46.2 +- 0.2	50.8 +- 0.1	54.5 +- 0.2	57.4 +- 0.2	60.0 +- 0.1	66.1 +- 0.0
s-DINO	19.8 +- 0.1	27.9 +- 0.3	36.7 +- 0.2	44.5 +- 0.2	51.3 +- 0.3	56.2 +- 0.1	60.4 +- 0.1	63.9 +- 0.1	71.1 +- 0.0
s-SwAV	17.2 +- 0.1	24.8 +- 0.4	33.6 +- 0.2	41.8 +- 0.3	49.1 +- 0.3	54.5 +- 0.1	59.0 +- 0.1	62.6 +- 0.1	70.2 +- 0.0
s-BarlowTwins	18.6 +- 0.2	26.2 +- 0.3	34.6 +- 0.3	42.1 +- 0.2	48.6 +- 0.2	53.7 +- 0.1	58.0 +- 0.1	61.6 +- 0.2	69.2 +- 0.0
s-OBoW	17.4 +- 0.1	24.2 +- 0.4	32.4 +- 0.2	39.7 +- 0.2	46.4 +- 0.2	51.8 +- 0.1	56.5 +- 0.1	60.4 +- 0.1	68.6 +- 0.0
s-BYOL	20.5 +- 0.5	28.3 +- 0.3	36.9 +- 0.2	44.7 +- 0.1	50.6 +- 0.2	55.6 +- 0.2	59.5 +- 0.1	62.8 +- 0.1	69.7 +- 0.0
s-SimCLR-v2	15.7 +- 0.2	22.7 +- 0.3	30.8 +- 0.1	38.6 +- 0.2	45.6 +- 0.2	51.1 +- 0.2	55.7 +- 0.1	59.8 +- 0.1	68.2 +- 0.0
s-MoCo-v2	19.7 +- 0.3	25.9 +- 0.3	32.6 +- 0.3	39.0 +- 0.2	45.0 +- 0.1	50.0 +- 0.2	54.4 +- 0.1	58.3 +- 0.1	67.1 +- 0.0
s-MoCHi	21.0 +- 0.2	26.9 +- 0.3	33.7 +- 0.3	39.2 +- 0.3	44.7 +- 0.1	49.3 +- 0.2	53.5 +- 0.1	57.3 +- 0.1	65.8 +- 0.0
s-CompReSS	21.6 +- 0.2	28.8 +- 0.4	36.2 +- 0.2	42.3 +- 0.2	47.8 +- 0.2	52.0 +- 0.1	55.7 +- 0.1	58.8 +- 0.1	65.9 +- 0.0
s-InfoMin	21.4 +- 0.1	27.7 +- 0.3	34.4 +- 0.4	40.6 +- 0.3	46.4 +- 0.1	51.2 +- 0.1	55.4 +- 0.2	59.2 +- 0.1	67.9 +- 0.1
d-Semi-Sup.	28.1 +- 0.4	37.6 +- 0.3	46.1 +- 0.2	52.6 +- 0.1	57.6 +- 0.2	61.1 +- 0.1	64.1 +- 0.2	66.5 +- 0.1	71.8 +- 0.0
d-Semi-Weakly-Sup.	30.0 +- 0.3	39.4 +- 0.4	47.9 +- 0.2	54.4 +- 0.1	59.2 +- 0.1	62.7 +- 0.1	65.6 +- 0.1	68.0 +- 0.1	72.9 +- 0.0
d-MoPro	26.0 +- 0.2	34.5 +- 0.4	42.0 +- 0.2	47.8 +- 0.2	52.6 +- 0.1	56.2 +- 0.1	59.2 +- 0.1	61.7 +- 0.2	68.0 +- 0.1
d-CLIP	22.7 +- 0.8	31.4 +- 0.4	40.3 +- 0.2	48.0 +- 0.3	53.7 +- 0.2	58.2 +- 0.2	61.7 +- 0.1	64.4 +- 0.1	69.8 +- 0.1
r-ReLabel	30.3 +- 0.2	37.1 +- 0.4	43.5 +- 0.3	48.1 +- 0.3	52.1 +- 0.1	55.2 +- 0.2	57.8 +- 0.1	59.9 +- 0.1	64.9 +- 0.1
r-CutMix	29.1 +- 0.1	35.8 +- 0.3	41.9 +- 0.1	46.8 +- 0.2	51.0 +- 0.1	54.1 +- 0.1	56.9 +- 0.1	59.1 +- 0.1	64.8 +- 0.1
r-Mixup	28.4 +- 0.3	35.9 +- 0.4	42.5 +- 0.3	47.8 +- 0.3	52.1 +- 0.1	55.1 +- 0.1	57.8 +- 0.1	60.0 +- 0.2	65.0 +- 0.1
r-Manifold Mixup	28.4 +- 0.3	35.6 +- 0.3	42.3 +- 0.2	47.6 +- 0.3	52.0 +- 0.2	55.0 +- 0.2	57.9 +- 0.2	60.2 +- 0.1	65.5 +- 0.1
r-AdvRobust	25.2 +- 0.4	33.5 +- 0.3	41.4 +- 0.1	47.6 +- 0.2	52.6 +- 0.1	56.2 +- 0.2	59.0 +- 0.1	61.4 +- 0.1	67.4 +- 0.0
r-MEAL-v2	30.8 +- 0.3	38.5 +- 0.3	45.6 +- 0.3	51.0 +- 0.2	55.7 +- 0.2	59.0 +- 0.1	61.9 +- 0.1	64.3 +- 0.1	69.8 +- 0.1

Table E: **Top-1 accuracies obtained by linear classifiers on L_1.** Table view corresponding to the 2nd row in Fig. F.

	N-shots								
Model	2^0	2^1	2^2	2^3	2^4	2^5	2^6	2^7	All
ResNet50	18.7 +- 0.4	26.2 +- 0.3	33.8 +- 0.3	40.4 +- 0.1	45.6 +- 0.3	49.9 +- 0.2	53.2 +- 0.3	56.0 +- 0.2	63.1 +- 0.0
a-ResNet-152	20.5 +- 0.4	28.6 +- 0.2	36.1 +- 0.2	42.8 +- 0.1	48.1 +- 0.2	52.2 +- 0.2	55.4 +- 0.2	58.1 +- 0.1	65.0 +- 0.0
a-T2T-ViTt-14	22.4 +- 0.4	29.6 +- 0.4	37.1 +- 0.3	43.4 +- 0.2	48.3 +- 0.3	52.4 +- 0.2	55.8 +- 0.2	58.6 +- 0.1	63.9 +- 0.0
a-DeiT-S	21.6 +- 0.4	28.9 +- 0.4	36.3 +- 0.4	42.7 +- 0.1	47.7 +- 0.2	51.8 +- 0.1	55.1 +- 0.1	57.9 +- 0.1	63.4 +- 0.0
a-DeiT-S distilled	23.1 +- 0.3	31.3 +- 0.4	39.2 +- 0.3	45.8 +- 0.2	50.6 +- 0.2	54.5 +- 0.2	57.7 +- 0.1	60.1 +- 0.1	65.1 +- 0.0
a-DeiT-B distilled	24.1 +- 0.3	32.1 +- 0.3	39.9 +- 0.1	47.0 +- 0.0	51.9 +- 0.2	56.2 +- 0.2	59.5 +- 0.1	62.2 +- 0.1	68.1 +- 0.0
a-Inception-v3	19.2 +- 0.3	26.1 +- 0.4	33.0 +- 0.3	39.1 +- 0.1	44.1 +- 0.2	48.3 +- 0.2	51.7 +- 0.3	54.7 +- 0.1	61.8 +- 0.1
a-EfficientNet-B1	18.1 +- 0.4	26.0 +- 0.2	34.2 +- 0.2	41.6 +- 0.2	47.2 +- 0.1	51.6 +- 0.2	55.1 +- 0.1	57.9 +- 0.1	64.2 +- 0.1
a-EfficientNet-B4	20.0 +- 0.3	27.9 +- 0.2	36.9 +- 0.3	44.3 +- 0.3	50.2 +- 0.3	54.5 +- 0.1	57.8 +- 0.1	60.6 +- 0.2	66.8 +- 0.0
a-NAT-M4	24.0 +- 0.4	32.9 +- 0.2	40.7 +- 0.4	46.9 +- 0.1	51.8 +- 0.1	55.6 +- 0.1	58.7 +- 0.1	61.3 +- 0.1	67.1 +- 0.0
a-VGG19	18.1 +- 0.5	25.0 +- 0.2	31.8 +- 0.0	38.1 +- 0.2	43.0 +- 0.3	47.3 +- 0.3	50.8 +- 0.2	53.7 +- 0.3	61.0 +- 0.0
s-DINO	15.6 +- 0.4	22.6 +- 0.2	30.6 +- 0.3	38.3 +- 0.2	45.3 +- 0.2	50.9 +- 0.1	55.3 +- 0.1	59.2 +- 0.1	67.2 +- 0.0
s-SwAV	13.5 +- 0.4	19.7 +- 0.3	27.4 +- 0.3	35.5 +- 0.1	42.8 +- 0.2	48.8 +- 0.2	53.7 +- 0.1	57.9 +- 0.1	66.5 +- 0.0
s-BarlowTwins	14.2 +- 0.3	20.7 +- 0.2	28.4 +- 0.2	36.1 +- 0.2	42.9 +- 0.1	48.3 +- 0.1	53.0 +- 0.2	57.0 +- 0.1	65.3 +- 0.0
s-OBoW	12.8 +- 0.5	18.4 +- 0.2	25.2 +- 0.3	32.4 +- 0.2	39.3 +- 0.2	45.3 +- 0.2	50.6 +- 0.1	54.9 +- 0.1	64.2 +- 0.1
s-BYOL	15.5 +- 0.3	22.4 +- 0.4	30.3 +- 0.3	37.8 +- 0.1	44.5 +- 0.2	49.7 +- 0.3	54.1 +- 0.2	58.0 +- 0.2	65.6 +- 0.0
s-SimCLR-v2	12.5 +- 0.2	18.3 +- 0.3	25.3 +- 0.3	32.9 +- 0.3	39.9 +- 0.1	45.8 +- 0.2	50.8 +- 0.1	55.1 +- 0.1	63.7 +- 0.0
s-MoCo-v2	13.4 +- 0.4	18.8 +- 0.3	25.2 +- 0.2	31.7 +- 0.1	38.3 +- 0.3	43.9 +- 0.2	48.9 +- 0.2	53.2 +- 0.1	62.6 +- 0.1
s-MoCHi	13.7 +- 0.2	18.7 +- 0.3	24.9 +- 0.4	31.0 +- 0.2	37.0 +- 0.1	42.5 +- 0.3	47.3 +- 0.2	51.6 +- 0.1	61.3 +- 0.0
s-CompReSS	15.1 +- 0.4	21.3 +- 0.4	28.2 +- 0.2	34.7 +- 0.2	40.5 +- 0.2	45.4 +- 0.2	49.7 +- 0.1	53.4 +- 0.1	61.3 +- 0.0
s-InfoMin	14.1 +- 0.4	19.6 +- 0.4	26.0 +- 0.2	32.4 +- 0.1	38.7 +- 0.2	44.4 +- 0.2	49.4 +- 0.2	53.7 +- 0.1	63.1 +- 0.0
d-Semi-Sup.	21.7 +- 0.4	30.2 +- 0.4	38.4 +- 0.3	45.4 +- 0.1	50.9 +- 0.2	55.1 +- 0.1	58.6 +- 0.2	61.4 +- 0.2	67.5 +- 0.0
d-Semi-Weakly-Sup.	23.0 +- 0.5	31.7 +- 0.6	40.1 +- 0.2	47.1 +- 0.2	52.8 +- 0.2	56.8 +- 0.2	60.3 +- 0.2	63.0 +- 0.1	68.8 +- 0.0
d-MoPro	18.7 +- 0.4	25.9 +- 0.4	33.2 +- 0.4	39.7 +- 0.1	45.0 +- 0.2	49.2 +- 0.2	52.7 +- 0.3	55.7 +- 0.2	63.2 +- 0.2
d-CLIP	19.4 +- 0.5	27.4 +- 0.4	35.6 +- 0.3	43.2 +- 0.2	49.1 +- 0.3	53.7 +- 0.1	57.6 +- 0.1	60.5 +- 0.1	66.4 +- 0.0
r-ReLabel	20.0 +- 0.2	26.6 +- 0.4	33.1 +- 0.2	38.6 +- 0.1	43.0 +- 0.2	46.8 +- 0.1	49.8 +- 0.2	52.4 +- 0.1	58.8 +- 0.1
r-CutMix	19.0 +- 0.2	25.1 +- 0.4	31.6 +- 0.3	37.1 +- 0.1	41.8 +- 0.2	45.8 +- 0.1	49.1 +- 0.1	51.8 +- 0.0	58.5 +- 0.1
r-Mixup	18.9 +- 0.3	25.7 +- 0.4	32.5 +- 0.3	38.4 +- 0.1	43.2 +- 0.1	47.1 +- 0.1	50.3 +- 0.1	53.0 +- 0.2	59.5 +- 0.1
r-Manifold Mixup	19.0 +- 0.3	25.8 +- 0.2	32.5 +- 0.3	38.5 +- 0.2	43.3 +- 0.2	47.3 +- 0.2	50.6 +- 0.1	53.4 +- 0.2	59.9 +- 0.1
r-AdvRobust	18.3 +- 0.5	25.5 +- 0.2	33.2 +- 0.3	39.9 +- 0.1	45.3 +- 0.2	49.4 +- 0.2	52.7 +- 0.2	55.5 +- 0.3	62.8 +- 0.0
r-MEAL-v2	20.9 +- 0.2	28.4 +- 0.4	35.6 +- 0.3	42.2 +- 0.1	47.3 +- 0.2	51.4 +- 0.3	55.1 +- 0.3	58.0 +- 0.2	64.6 +- 0.1

Table F: **Top-1 accuracies obtained by linear classifiers on L_2.** Table view corresponding to the 3rd row in Fig. F.

	N-shots								
Model	2^0	2^1	2^2	2^3	2^4	2^5	2^6	2^7	All
ResNet50	16.7 +- 0.5	23.9 +- 0.2	30.7 +- 0.1	37.0 +- 0.1	42.0 +- 0.1	45.9 +- 0.1	49.1 +- 0.1	51.9 +- 0.2	59.0 +- 0.0
a-ResNet-152	18.1 +- 0.4	25.5 +- 0.4	32.7 +- 0.2	38.8 +- 0.2	43.8 +- 0.1	47.7 +- 0.2	50.8 +- 0.1	53.4 +- 0.1	60.5 +- 0.0
a-T2T-ViTt-14	19.3 +- 0.4	26.5 +- 0.4	33.3 +- 0.2	39.4 +- 0.1	44.4 +- 0.2	48.2 +- 0.1	51.6 +- 0.1	54.3 +- 0.1	59.7 +- 0.0
a-DeiT-S	18.4 +- 0.6	25.9 +- 0.3	32.9 +- 0.4	38.8 +- 0.2	43.8 +- 0.3	47.8 +- 0.1	51.0 +- 0.1	53.7 +- 0.1	59.1 +- 0.0
a-DeiT-S distilled	19.9 +- 0.5	27.8 +- 0.3	35.2 +- 0.3	41.4 +- 0.2	46.5 +- 0.2	50.2 +- 0.1	53.1 +- 0.1	55.8 +- 0.1	60.7 +- 0.0
a-DeiT-B distilled	21.7 +- 0.5	29.1 +- 0.2	36.6 +- 0.2	43.0 +- 0.1	48.1 +- 0.1	52.2 +- 0.1	55.6 +- 0.1	58.3 +- 0.1	64.4 +- 0.1
a-Inception-v3	16.5 +- 0.2	22.9 +- 0.2	29.2 +- 0.3	35.2 +- 0.2	40.1 +- 0.1	44.1 +- 0.1	47.5 +- 0.1	50.5 +- 0.1	57.2 +- 0.1
a-EfficientNet-B1	16.2 +- 0.5	23.4 +- 0.2	31.3 +- 0.2	37.9 +- 0.3	43.6 +- 0.2	47.6 +- 0.2	51.0 +- 0.1	53.9 +- 0.1	60.2 +- 0.1
a-EfficientNet-B4	17.8 +- 0.3	25.3 +- 0.3	33.8 +- 0.2	40.7 +- 0.2	46.2 +- 0.2	50.4 +- 0.2	53.6 +- 0.1	56.4 +- 0.2	62.7 +- 0.1
a-NAT-M4	21.3 +- 0.3	29.0 +- 0.3	36.6 +- 0.2	42.5 +- 0.2	47.5 +- 0.1	51.2 +- 0.1	54.3 +- 0.0	56.9 +- 0.1	62.8 +- 0.0
a-VGG19	16.0 +- 0.4	22.1 +- 0.4	28.8 +- 0.4	34.6 +- 0.3	39.3 +- 0.2	43.4 +- 0.2	46.7 +- 0.1	49.9 +- 0.2	57.0 +- 0.1
s-DINO	14.7 +- 0.5	21.9 +- 0.3	29.4 +- 0.2	36.6 +- 0.3	42.9 +- 0.1	48.0 +- 0.2	52.3 +- 0.1	55.8 +- 0.1	63.2 +- 0.0
s-SwAV	12.9 +- 0.5	19.4 +- 0.2	26.8 +- 0.2	34.3 +- 0.2	40.9 +- 0.1	46.3 +- 0.2	51.0 +- 0.1	54.7 +- 0.1	62.5 +- 0.0
s-BarlowTwins	13.2 +- 0.5	19.6 +- 0.3	26.8 +- 0.2	33.8 +- 0.2	40.0 +- 0.1	45.4 +- 0.1	49.6 +- 0.1	53.3 +- 0.1	61.3 +- 0.0
s-OBoW	11.8 +- 0.3	17.6 +- 0.2	23.9 +- 0.1	30.9 +- 0.3	37.2 +- 0.1	42.8 +- 0.2	47.6 +- 0.1	51.6 +- 0.2	60.5 +- 0.1
s-BYOL	14.4 +- 0.5	21.0 +- 0.3	28.4 +- 0.2	35.4 +- 0.3	41.6 +- 0.1	46.5 +- 0.1	50.7 +- 0.1	54.1 +- 0.1	61.5 +- 0.0
s-SimCLR-v2	11.8 +- 0.4	17.9 +- 0.1	24.7 +- 0.1	31.5 +- 0.2	37.9 +- 0.2	43.2 +- 0.1	48.0 +- 0.2	52.0 +- 0.1	59.9 +- 0.1
s-MoCo-v2	12.5 +- 0.4	17.9 +- 0.2	24.2 +- 0.1	30.3 +- 0.3	36.4 +- 0.2	41.7 +- 0.1	46.2 +- 0.1	50.3 +- 0.2	59.0 +- 0.0
s-MoCHi	12.6 +- 0.5	17.9 +- 0.3	23.6 +- 0.1	29.4 +- 0.2	35.2 +- 0.1	40.2 +- 0.1	44.7 +- 0.1	48.7 +- 0.1	57.5 +- 0.1
s-CompReSS	13.8 +- 0.4	20.1 +- 0.3	26.4 +- 0.4	32.7 +- 0.1	38.3 +- 0.2	42.9 +- 0.2	46.8 +- 0.2	50.3 +- 0.1	57.7 +- 0.0
s-InfoMin	12.9 +- 0.5	18.5 +- 0.3	24.7 +- 0.3	30.8 +- 0.2	36.8 +- 0.2	42.0 +- 0.2	46.6 +- 0.2	50.7 +- 0.2	59.5 +- 0.1
d-Semi-Sup.	19.5 +- 0.5	27.5 +- 0.2	35.1 +- 0.2	41.6 +- 0.3	46.8 +- 0.2	51.0 +- 0.1	54.3 +- 0.2	57.3 +- 0.1	63.3 +- 0.0
d-Semi-Weakly-Sup.	20.2 +- 0.4	28.6 +- 0.3	36.4 +- 0.1	43.3 +- 0.2	48.5 +- 0.1	52.6 +- 0.1	55.8 +- 0.1	58.7 +- 0.1	64.5 +- 0.0
d-MoPro	16.5 +- 0.4	23.8 +- 0.3	30.4 +- 0.3	37.0 +- 0.1	41.9 +- 0.1	45.9 +- 0.1	49.3 +- 0.0	52.3 +- 0.2	59.6 +- 0.1
d-CLIP	19.0 +- 0.3	26.7 +- 0.4	34.8 +- 0.4	41.9 +- 0.2	47.4 +- 0.2	51.7 +- 0.1	55.3 +- 0.2	58.0 +- 0.1	63.5 +- 0.0
r-ReLabel	17.0 +- 0.4	23.5 +- 0.4	29.6 +- 0.2	35.1 +- 0.3	39.5 +- 0.2	43.0 +- 0.2	46.1 +- 0.2	48.6 +- 0.1	54.8 +- 0.1
r-CutMix	16.2 +- 0.3	22.3 +- 0.4	28.4 +- 0.2	33.8 +- 0.1	38.4 +- 0.2	42.1 +- 0.0	45.3 +- 0.1	48.0 +- 0.1	54.7 +- 0.1
r-Mixup	16.5 +- 0.5	23.1 +- 0.2	29.3 +- 0.2	35.1 +- 0.2	39.9 +- 0.1	43.5 +- 0.2	46.7 +- 0.1	49.3 +- 0.1	55.6 +- 0.0
r-Manifold Mixup	16.6 +- 0.4	23.0 +- 0.3	29.4 +- 0.1	35.2 +- 0.1	40.0 +- 0.1	43.8 +- 0.2	47.0 +- 0.2	49.7 +- 0.2	56.1 +- 0.1
r-AdvRobust	16.2 +- 0.6	23.3 +- 0.3	30.0 +- 0.1	36.4 +- 0.1	41.5 +- 0.1	45.5 +- 0.1	48.8 +- 0.1	51.5 +- 0.2	58.5 +- 0.0
r-MEAL-v2	18.3 +- 0.4	25.6 +- 0.4	32.6 +- 0.2	38.6 +- 0.1	43.7 +- 0.1	47.6 +- 0.2	50.9 +- 0.2	53.9 +- 0.2	60.7 +- 0.0

Table G: **Top-1 accuracies obtained by linear classifiers on L_3.** Table view corresponding to the 4th row in Fig. F.

	N-shots								
Model	2^0	2^1	2^2	2^3	2^4	2^5	2^6	2^7	All
ResNet50	15.1 +- 0.4	21.7 +- 0.3	29.1 +- 0.3	35.0 +- 0.2	40.4 +- 0.0	44.5 +- 0.2	47.9 +- 0.1	50.9 +- 0.2	58.1 +- 0.0
a-ResNet-152	16.5 +- 0.5	23.2 +- 0.3	30.6 +- 0.2	36.8 +- 0.1	42.1 +- 0.1	46.1 +- 0.1	49.5 +- 0.1	52.3 +- 0.1	59.5 +- 0.0
a-T2T-ViTt-14	17.8 +- 0.3	24.7 +- 0.4	31.3 +- 0.2	37.2 +- 0.3	42.3 +- 0.1	46.8 +- 0.2	50.1 +- 0.1	52.9 +- 0.1	58.2 +- 0.1
a-DeiT-S	17.2 +- 0.3	23.9 +- 0.5	30.7 +- 0.2	36.8 +- 0.3	42.1 +- 0.1	46.5 +- 0.2	50.0 +- 0.2	52.7 +- 0.1	57.9 +- 0.0
a-DeiT-S distilled	18.7 +- 0.4	25.8 +- 0.3	33.4 +- 0.2	39.6 +- 0.1	44.6 +- 0.1	48.7 +- 0.1	52.0 +- 0.1	54.7 +- 0.1	59.9 +- 0.0
a-DeiT-B distilled	20.0 +- 0.2	27.2 +- 0.1	34.8 +- 0.2	41.4 +- 0.3	47.0 +- 0.2	51.3 +- 0.2	54.7 +- 0.1	57.5 +- 0.1	63.4 +- 0.1
a-Inception-v3	14.7 +- 0.3	20.4 +- 0.3	27.2 +- 0.3	32.7 +- 0.2	37.8 +- 0.2	41.8 +- 0.2	45.4 +- 0.2	48.4 +- 0.1	55.6 +- 0.1
a-EfficientNet-B1	15.3 +- 0.2	21.9 +- 0.5	29.5 +- 0.2	36.5 +- 0.2	42.1 +- 0.1	46.5 +- 0.0	49.9 +- 0.1	52.7 +- 0.1	59.2 +- 0.1
a-EfficientNet-B4	16.4 +- 0.2	23.8 +- 0.4	31.9 +- 0.1	39.1 +- 0.1	45.0 +- 0.3	49.4 +- 0.1	52.8 +- 0.1	55.6 +- 0.2	62.0 +- 0.1
a-NAT-M4	19.6 +- 0.4	27.4 +- 0.4	35.1 +- 0.2	40.8 +- 0.1	45.7 +- 0.1	49.7 +- 0.1	52.9 +- 0.1	55.7 +- 0.1	61.6 +- 0.0
a-VGG19	14.2 +- 0.3	20.1 +- 0.2	26.4 +- 0.2	32.2 +- 0.1	37.2 +- 0.1	41.1 +- 0.2	44.6 +- 0.2	47.8 +- 0.2	54.7 +- 0.0
s-DINO	13.6 +- 0.4	20.2 +- 0.3	28.0 +- 0.1	35.0 +- 0.3	41.7 +- 0.1	46.9 +- 0.2	51.3 +- 0.1	54.9 +- 0.1	62.6 +- 0.0
s-SwAV	11.8 +- 0.4	18.0 +- 0.2	25.3 +- 0.2	32.5 +- 0.3	39.2 +- 0.2	44.8 +- 0.2	49.4 +- 0.1	53.4 +- 0.1	61.4 +- 0.0
s-BarlowTwins	12.2 +- 0.4	18.3 +- 0.5	25.8 +- 0.2	32.7 +- 0.3	39.3 +- 0.2	44.6 +- 0.2	49.0 +- 0.1	52.7 +- 0.2	60.4 +- 0.0
s-OBoW	11.0 +- 0.3	16.2 +- 0.4	22.7 +- 0.2	29.2 +- 0.3	35.6 +- 0.2	41.1 +- 0.2	46.0 +- 0.2	50.1 +- 0.1	59.0 +- 0.0
s-BYOL	13.1 +- 0.3	19.4 +- 0.3	26.8 +- 0.2	33.8 +- 0.2	40.1 +- 0.2	45.1 +- 0.2	49.5 +- 0.1	53.0 +- 0.2	60.4 +- 0.0
s-SimCLR-v2	11.0 +- 0.3	16.5 +- 0.3	23.3 +- 0.1	30.0 +- 0.3	36.4 +- 0.1	42.0 +- 0.1	46.9 +- 0.1	50.7 +- 0.1	58.9 +- 0.0
s-MoCo-v2	11.6 +- 0.2	16.6 +- 0.3	22.9 +- 0.3	28.9 +- 0.3	34.8 +- 0.3	40.2 +- 0.2	44.8 +- 0.3	48.8 +- 0.1	57.5 +- 0.0
s-MoCHi	11.4 +- 0.2	16.4 +- 0.2	22.0 +- 0.2	27.8 +- 0.2	33.4 +- 0.2	38.3 +- 0.2	43.1 +- 0.2	47.1 +- 0.1	56.0 +- 0.1
s-CompReSS	12.7 +- 0.3	18.5 +- 0.3	25.2 +- 0.2	31.2 +- 0.2	36.9 +- 0.2	41.5 +- 0.2	45.7 +- 0.2	49.1 +- 0.1	56.7 +- 0.0
s-InfoMin	11.9 +- 0.3	17.0 +- 0.3	23.2 +- 0.2	29.3 +- 0.2	35.3 +- 0.1	40.6 +- 0.1	45.4 +- 0.3	49.5 +- 0.2	58.7 +- 0.1
d-Semi-Sup.	18.3 +- 0.4	25.9 +- 0.3	34.2 +- 0.2	40.8 +- 0.3	46.0 +- 0.1	50.2 +- 0.2	53.7 +- 0.1	56.5 +- 0.2	62.6 +- 0.0
d-Semi-Weakly-Sup.	19.2 +- 0.5	27.1 +- 0.4	35.3 +- 0.2	41.9 +- 0.2	47.3 +- 0.2	51.5 +- 0.2	55.0 +- 0.1	57.7 +- 0.1	63.4 +- 0.0
d-MoPro	15.6 +- 0.3	22.4 +- 0.2	29.5 +- 0.3	35.6 +- 0.2	41.0 +- 0.1	45.2 +- 0.2	48.6 +- 0.1	51.5 +- 0.1	59.0 +- 0.2
d-CLIP	18.6 +- 0.4	26.8 +- 0.3	35.3 +- 0.3	42.5 +- 0.3	48.2 +- 0.2	52.7 +- 0.2	56.1 +- 0.1	58.7 +- 0.1	64.0 +- 0.0
r-ReLabel	15.5 +- 0.4	21.3 +- 0.3	27.6 +- 0.1	32.7 +- 0.1	37.2 +- 0.1	41.1 +- 0.1	44.1 +- 0.1	46.7 +- 0.1	53.2 +- 0.0
r-CutMix	14.6 +- 0.3	20.1 +- 0.4	26.1 +- 0.2	31.3 +- 0.4	36.1 +- 0.1	40.0 +- 0.1	43.4 +- 0.1	46.2 +- 0.2	52.9 +- 0.1
r-Mixup	15.1 +- 0.4	20.7 +- 0.4	27.3 +- 0.0	32.8 +- 0.3	37.7 +- 0.1	41.7 +- 0.3	45.0 +- 0.1	47.5 +- 0.0	53.7 +- 0.1
r-Manifold Mixup	15.6 +- 0.0	20.9 +- 0.4	27.4 +- 0.2	32.9 +- 0.2	37.8 +- 0.1	41.9 +- 0.2	45.2 +- 0.2	47.9 +- 0.2	54.3 +- 0.1
r-AdvRobust	14.8 +- 0.4	21.3 +- 0.4	28.5 +- 0.1	34.5 +- 0.2	39.7 +- 0.2	44.0 +- 0.2	47.3 +- 0.2	50.3 +- 0.2	57.5 +- 0.0
r-MEAL-v2	16.5 +- 0.4	23.3 +- 0.4	30.4 +- 0.2	36.5 +- 0.2	41.7 +- 0.1	45.8 +- 0.2	49.4 +- 0.1	52.4 +- 0.1	59.4 +- 0.0

Table H: **Top-1 accuracies obtained by linear classifiers on L_4.** Table view corresponding to the 5th row in Fig. F.

	N-shots								
Model	2^0	2^1	2^2	2^3	2^4	2^5	2^6	2^7	All
ResNet50	12.2 +- 0.1	17.6 +- 0.3	23.4 +- 0.1	29.2 +- 0.2	33.8 +- 0.2	38.0 +- 0.1	41.5 +- 0.1	44.4 +- 0.1	52.0 +- 0.0
a-ResNet-152	13.1 +- 0.2	18.7 +- 0.3	24.5 +- 0.2	30.3 +- 0.3	34.7 +- 0.2	39.0 +- 0.2	42.4 +- 0.2	45.3 +- 0.1	52.8 +- 0.0
a-T2T-ViTt-14	14.2 +- 0.2	19.2 +- 0.4	25.4 +- 0.2	31.0 +- 0.2	36.0 +- 0.1	40.2 +- 0.1	43.6 +- 0.2	46.5 +- 0.1	51.9 +- 0.0
a-DeiT-S	13.9 +- 0.3	18.8 +- 0.4	25.0 +- 0.1	30.6 +- 0.1	35.5 +- 0.2	39.8 +- 0.1	43.2 +- 0.1	46.0 +- 0.1	51.5 +- 0.0
a-DeiT-S distilled	15.0 +- 0.4	20.6 +- 0.3	27.1 +- 0.1	32.7 +- 0.1	37.6 +- 0.2	41.6 +- 0.1	44.9 +- 0.2	47.5 +- 0.2	52.8 +- 0.1
a-DeiT-B distilled	16.0 +- 0.3	22.1 +- 0.4	29.1 +- 0.2	35.4 +- 0.1	40.3 +- 0.3	44.8 +- 0.1	48.4 +- 0.0	51.5 +- 0.1	57.5 +- 0.0
a-Inception-v3	11.5 +- 0.3	16.4 +- 0.2	21.4 +- 0.3	26.8 +- 0.2	31.1 +- 0.1	35.3 +- 0.2	38.8 +- 0.2	41.9 +- 0.1	48.7 +- 0.1
a-EfficientNet-B1	12.4 +- 0.1	18.3 +- 0.2	24.9 +- 0.3	30.8 +- 0.4	36.1 +- 0.2	40.4 +- 0.1	43.8 +- 0.2	46.8 +- 0.1	52.9 +- 0.0
a-EfficientNet-B4	13.4 +- 0.2	19.7 +- 0.5	26.8 +- 0.2	33.3 +- 0.2	38.6 +- 0.2	42.9 +- 0.2	46.3 +- 0.2	49.2 +- 0.1	55.9 +- 0.0
a-NAT-M4	15.3 +- 0.3	21.6 +- 0.2	28.0 +- 0.1	33.6 +- 0.2	38.5 +- 0.2	42.6 +- 0.2	45.8 +- 0.1	48.5 +- 0.1	54.6 +- 0.1
a-VGG19	11.0 +- 0.2	15.7 +- 0.3	21.1 +- 0.3	25.8 +- 0.2	30.4 +- 0.2	34.4 +- 0.2	37.8 +- 0.2	40.9 +- 0.1	47.7 +- 0.1
s-DINO	12.6 +- 0.1	18.6 +- 0.3	25.3 +- 0.1	32.3 +- 0.3	38.2 +- 0.1	43.2 +- 0.1	47.3 +- 0.1	50.8 +- 0.1	57.6 +- 0.0
s-SwAV	10.7 +- 0.1	16.1 +- 0.3	22.7 +- 0.1	29.7 +- 0.3	35.9 +- 0.2	41.2 +- 0.2	45.6 +- 0.1	49.4 +- 0.0	56.4 +- 0.0
s-BarlowTwins	10.9 +- 0.2	16.3 +- 0.3	22.6 +- 0.2	29.5 +- 0.2	35.3 +- 0.2	40.4 +- 0.2	44.8 +- 0.1	48.4 +- 0.2	55.4 +- 0.0
s-OBoW	9.4 +- 0.2	14.0 +- 0.1	19.6 +- 0.2	25.9 +- 0.2	31.9 +- 0.2	37.2 +- 0.1	41.9 +- 0.2	45.9 +- 0.1	54.2 +- 0.0
s-BYOL	11.5 +- 0.2	16.9 +- 0.3	23.4 +- 0.1	30.4 +- 0.2	35.8 +- 0.1	40.9 +- 0.1	45.1 +- 0.1	48.5 +- 0.1	55.2 +- 0.0
s-SimCLR-v2	9.5 +- 0.2	14.3 +- 0.2	20.0 +- 0.2	26.8 +- 0.3	32.3 +- 0.2	37.7 +- 0.2	42.5 +- 0.2	46.3 +- 0.1	53.7 +- 0.0
s-MoCo-v2	9.7 +- 0.3	14.1 +- 0.3	19.4 +- 0.2	25.3 +- 0.2	30.8 +- 0.2	36.0 +- 0.3	40.6 +- 0.2	44.6 +- 0.1	52.3 +- 0.0
s-MoCHi	9.4 +- 0.2	13.5 +- 0.4	18.2 +- 0.3	23.8 +- 0.1	28.8 +- 0.3	33.9 +- 0.1	38.3 +- 0.2	42.3 +- 0.2	50.5 +- 0.1
s-CompReSS	11.1 +- 0.2	16.0 +- 0.2	21.8 +- 0.2	27.9 +- 0.2	32.9 +- 0.2	37.4 +- 0.2	41.3 +- 0.2	44.7 +- 0.1	51.7 +- 0.0
s-InfoMin	9.9 +- 0.2	14.2 +- 0.1	19.3 +- 0.3	25.3 +- 0.2	30.8 +- 0.1	36.2 +- 0.1	40.9 +- 0.1	45.0 +- 0.1	52.9 +- 0.0
d-Semi-Sup.	15.2 +- 0.2	21.5 +- 0.3	28.5 +- 0.1	34.8 +- 0.2	39.5 +- 0.2	43.9 +- 0.1	47.5 +- 0.1	50.4 +- 0.1	56.3 +- 0.1
d-Semi-Weakly-Sup.	15.5 +- 0.1	21.8 +- 0.4	28.8 +- 0.3	35.2 +- 0.2	40.2 +- 0.2	44.5 +- 0.2	47.9 +- 0.1	50.9 +- 0.1	56.7 +- 0.0
d-MoPro	13.0 +- 0.1	18.7 +- 0.3	24.9 +- 0.2	30.9 +- 0.2	35.6 +- 0.1	39.8 +- 0.1	43.3 +- 0.1	46.4 +- 0.2	53.8 +- 0.1
d-CLIP	15.9 +- 0.1	22.3 +- 0.3	29.7 +- 0.5	36.1 +- 0.5	41.5 +- 0.2	45.6 +- 0.1	48.9 +- 0.1	51.5 +- 0.1	56.7 +- 0.0
r-ReLabel	12.0 +- 0.2	16.8 +- 0.4	21.6 +- 0.2	26.5 +- 0.3	30.4 +- 0.1	33.9 +- 0.1	37.0 +- 0.2	39.7 +- 0.1	46.4 +- 0.1
r-CutMix	11.4 +- 0.1	15.9 +- 0.2	20.6 +- 0.1	25.3 +- 0.2	29.3 +- 0.2	33.1 +- 0.1	36.4 +- 0.2	39.2 +- 0.1	45.8 +- 0.0
r-Mixup	11.8 +- 0.2	16.6 +- 0.2	21.7 +- 0.1	26.9 +- 0.2	31.0 +- 0.2	35.0 +- 0.1	38.3 +- 0.0	41.1 +- 0.1	47.3 +- 0.1
r-Manifold Mixup	11.8 +- 0.2	16.7 +- 0.2	21.9 +- 0.1	27.0 +- 0.1	31.2 +- 0.1	35.2 +- 0.1	38.5 +- 0.2	41.2 +- 0.2	47.9 +- 0.1
r-AdvRobust	11.8 +- 0.1	17.1 +- 0.4	23.0 +- 0.2	28.8 +- 0.1	33.4 +- 0.3	37.6 +- 0.1	41.0 +- 0.1	44.0 +- 0.2	51.6 +- 0.1
r-MEAL-v2	13.1 +- 0.2	18.5 +- 0.3	24.2 +- 0.2	29.9 +- 0.2	34.7 +- 0.3	38.9 +- 0.3	42.5 +- 0.3	45.6 +- 0.3	52.2 +- 0.0

Table I: **Top-1 accuracies obtained by linear classifiers on L_5.** Table view corresponding to the last row in Fig. F.