

Appendix A: *Few-Class Arena: A Benchmark for Efficient Selection of Vision Models and Dataset Difficulty Measurement: Supplementary Materials (Submission Number: 14)*

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: We claim to provide a benchmark tool in the *Few-Class Regime* and insights based on our comprehensive experiments. We release our benchmark tool on a github repository with a link provided (<https://github.com/fewclassarena/fca>) in the supplemental materials and in the abstract. We summarize our new insights in Section 1 Introduction and Section 4 Experimental Results in the main paper.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We list the limitations in **Limitations and Future Work** in the Section 5 Conclusion in the main paper as well as discussions in the supplemental materials.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper focuses on benchmarking existing models in the *Few-Class Regime* which doesn't include theoretical proof. We provide the intuition of our proposed Similarity-Based Silhouette Score in Section 3.6 Few-Class Similarity Benchmark (FC-Sim) and empirical results in 4.3 Results on FC-Sim in the main paper, as well as the extended details of mathematical notations and derivation in A.7 Extended Few-Class Similarity Benchmark (FC-Sim) Details in the supplemental materials.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental materials.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We detail the experimental settings in the A.3 Datasets, A.4 Model Training Details in the supplemental materials. Experiments can also be reproduced by the instructions in the github repository <https://github.com/fewclassarena/fca>.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may

be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We publicly release our source code in the github link: <https://github.com/fewclassarena/fca> with detailed instructions. We also attach a copy of the source code in the folder “fca” for reference. Users can follow the instructions to reproduce results. Users can also create the conda environment by the fca.yaml file in <https://github.com/fewclassarena/fca/blob/main/fca.yaml>. Pip packages with their versions are listed in <https://github.com/fewclassarena/fca/blob/main/requirements.txt>

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: We present the training and testing details in the A.3 Datasets and A.4 Model Training Details sections, including **Train/val splits**, **optimizers** and other hyperparameters (e.g., weight decay) in Table 4 in the supplemental materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[Yes\]](#)

Justification: Error bars are defined as standard deviation of Top-1 accuracies in 5 subsets for a specific N_{CL} , reported as confidence area (light blue or pink) in Fig. 1, 3, 4 and shaded area in Fig. 5. P-value is reported in Fig. 5 in the main paper, as well as Fig. 3-22 and P-values in Fig. 24, 25 in the supplemental materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide the details in A.9 Experiments Compute Resources in the supplemental materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: No human subjects are involved. Our few-class datasets are based on publicly available datasets. We use their licenses for our research.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss these in the Conclusion and Discussion (section 7) in the main paper.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: Our method poses no risk for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We mentioned in the main paper that the full list of models (with pre-trained models) and datasets is listed and cited in the supplemental materials, specifically Table 1 and 2 in A.1 Full Models on ImageNet and A.3 Datasets. We list all dataset licenses in Table 3 in A.3 Datasets. The use of MMPreTrain [1] has been properly cited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Our code to build this *Few-Class Arena* is documented in the README.md file from the repository <https://github.com/fewclassarena/fca>. We also attach a copy of the script in the “fca” folder.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and Research with Human Subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: No crowdsourcing or research with human subjects is involved in this paper.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No crowdsourcing or research with human subjects is involved in this paper. Therefore, no IRB approval is required in this work.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

340
341

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

342 A.1 Full Models on ImageNet

343 In practice, ImageNet serves as a common benchmark for vision neural networks. We list the details
 344 of 10 pre-trained models from MMPreTrain [1] in terms of Top-1 Accuracy and scale (Params) in
 Fig. 1.

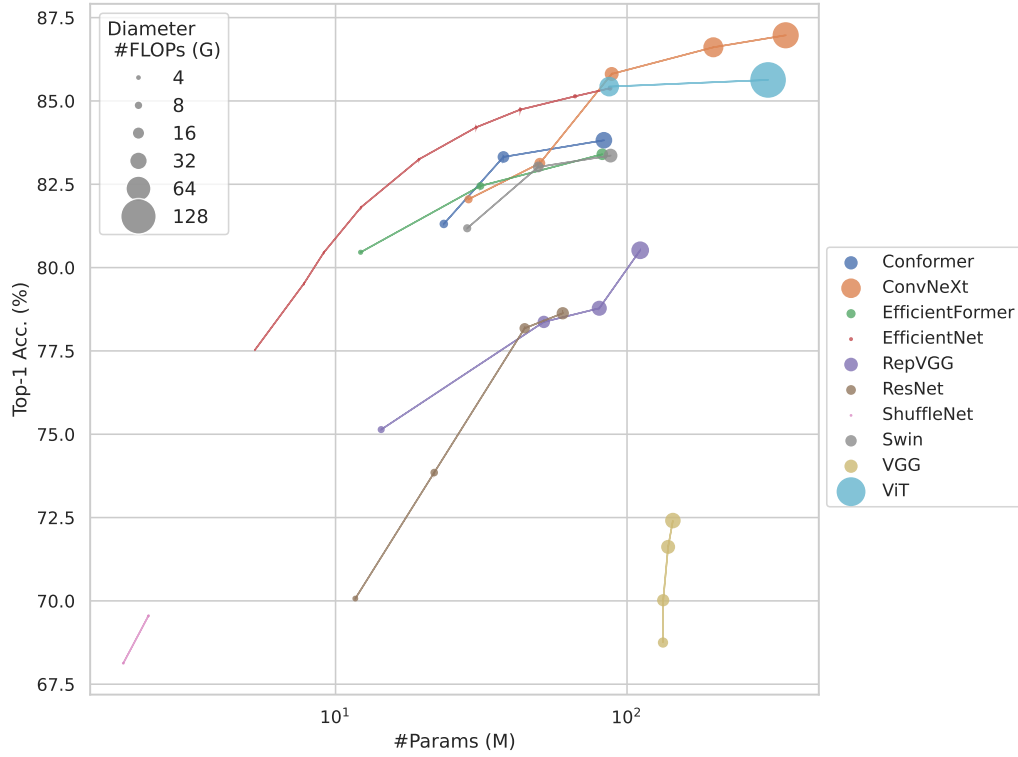


Figure 1: Top-1 Accuracy (%) vs. number of parameters and FLOPs (G) (size of circle) on ImageNet.

345

Model	Ref.	Model	Ref.
Conformer	[2]	ConvNeXt	[3]
EfficientFormer	[4]	EfficientNet	[5]
RepVGG	[6]	ResNet	[7]
ShuffleNet	[8]	Swin	[9]
VGG	[10]	ViT	[11]

Table 1: Full models pre-trained on ImageNet.

A.2 Extended Many-Class Full Dataset Trained Benchmark Results

A complete ranking of 10 models in 10 datasets is depicted in Fig. 2. Observe that the 10 models' rankings differ dramatically among 10 different datasets where each line changes from ImageNet1K (IN1K) to other datasets. This poses some questions whether rankings in existing benchmarks can be a reliable indicator for a practitioner to select an efficient neural network, especially when the deployed environment changes from application to application. A major variable in this process is the reduced number of classes from benchmark datasets to deployed environments in the *Few-Class Regime*. As such, our tool is developed to facilitate research in the *Few-Class Regime*.

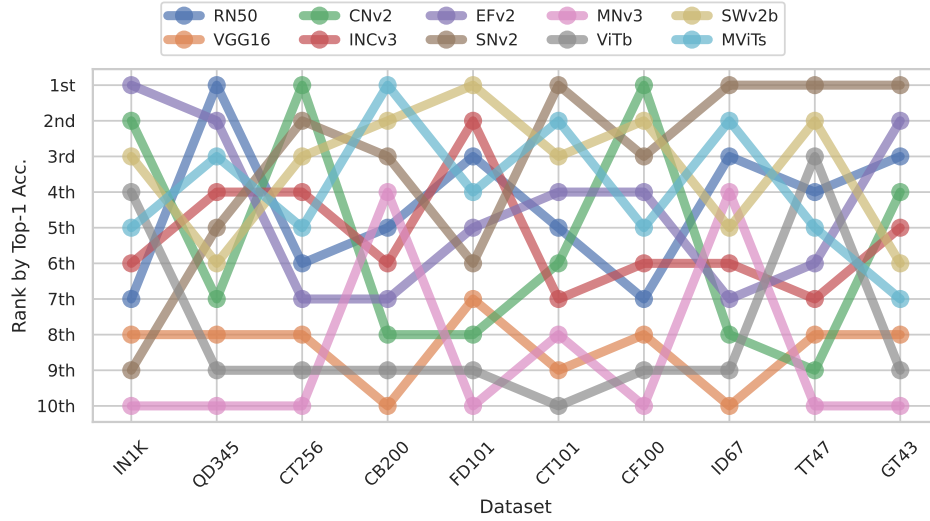


Figure 2: Extended Details of Fig. 2 (b) in the main paper. Full Ranking of 10 models across 10 datasets by Top-1 acc.

A.3 Datasets

Dataset information is presented in Table 2.

Dataset Name	Dataset Abbrev.	Ref.	Homepage	Path in FCA
Caltech 101	CT101	[12]	https://data.caltech.edu/records/mzrqj-6wc02	<code>tools/ncls/datasets/caltech101.py</code>
Caltech 256	CT256	[13]	https://data.caltech.edu/records/nyy15-4j048	<code>tools/ncls/datasets/caltech256.py</code>
CIFAR-100	CF100	[14]	https://www.cs.toronto.edu/~kriz/cifar.html https://github.com/knjcode/cifar2png	<code>tools/ncls/datasets/cifar100.py</code>
Caltech-UCSD Birds-200-2011	CB200	[15]	https://www.vision.caltech.edu/visipedia/CUB-200-2011.html https://data.caltech.edu/records/65de6-vp158/files/CUB_200_2011.tgz	<code>tools/ncls/datasets/cub200.py</code>
Food 101	FD101	[16]	https://vision.ee.ethz.ch/datasets_extra/food-101/ https://huggingface.co/datasets/food101	<code>tools/ncls/datasets/food101.py</code>
German Traffic Sign Recognition Benchmark	GT43	[17]	https://benchmark.ini.rub.de/	<code>tools/ncls/datasets/gtsrb43.py</code>
ImageNet Dataset	IN1K	[18]	https://www.image-net.org/challenges/LSVRC/2012/index.php	*
Indoor Scene Recognition	ID67	[19]	https://web.mit.edu/torralba/www/indoor.html	<code>tools/ncls/datasets/indoor67.py</code>
Quickdraw Dataset	QD345	[20]	https://github.com/googlecreativelab/quickdraw-dataset https://tensorflow.org/datasets/community_catalog/huggingface/quickdraw	<code>tools/ncls/datasets/quickdraw345.py</code>
Describable Textures Dataset	TT47	[21]	https://www.robots.ox.ac.uk/~vgg/data/dtd/index.html	<code>tools/ncls/datasets/textures47.py</code>

Table 2: Dataset information. * Note that ImageNet dataset format is used as the reference for other datasets. Therefore Path in FCA for ImageNet is not required.

License. We have searched available online resources and list the license of each dataset in Table 3. For licenses not found in the datasets or websites denoted as “*”, we assume they are non-commercial research use only.

Dataset	License	Dataset	License
CT101	CC BY 4.0	CT256	CC BY 4.0
CF100	*	CB200	*
FD101	*	GT43	*
IN1K	*	ID67	(DbCL) v1.0
QD345	CC BY 4.0	TT47	*

Table 3: Licenses of ten datasets.

358

Train/val splits. The dataset format follows the convention of ImageNet:

```

360 imagenet1k/
361     meta
362         train.txt
363         val.txt
364     train
365         <IMAGE_ID>.jpeg
366         ...
367     val
368         <IMAGE_ID>.jpeg
369         ...

```

where a .txt file stores a pair of image id and and class number in each row in the following format

```
<IMAGE_ID>.jpeg <CLASS_NUM>
```

We follow the same train/val splits when the original dataset has already provided. If the dataset does not have explicit splits, we first assign image IDs to all images, starting from 0, and select 4/5 of all images as training set and put the rest in the validation set. Specifially, when an image whose ID satisfies the condition $ID \% 5 == 0$, it will be moved to the validation set. Otherwise, it will be assigned as a training sample.

377 **A.4 Model Training Details**

Model training details are presented in Table 4.

Model	Model Abbrev.	Ref.	Optimizer	LR	Weight Decay	Other Params
ResNet50	RN50	[7]	SGD	0.1	0.0001	momentum=0.9
VGG16	VGG16	[22]	SGD	0.01	0.0001	momentum=0.9
ConvNeXt V2 Base	CNv2	[23]	AdamW	0.0025	0.05	eps=1e-8 betas=(0.9, 0.999)
Inception V3	INCv3	[24]	SGD	0.1	0.0001	momentum=0.9
EfficientNet V2 Medium	EFv2	[25]	SGD	4e-3	0.1	momentum=0.9 clip_grad: max_norm=5.0
ShuffleNet V2	SNv2	[26]	SGD	0.5	0.9	momentum=0.00004
MobileNet V3 Small	MNv3	[27]	RMSprop	0.064	1e-5	alpha=0.9 momentum=0.9 eps=0.0316
Vision Transformer Base	ViTb	[11]	AdamW	0.003	0.3	-
Swin Transformer V2 Base	SWv2b	[28]	AdamW	1e-4	0.05	eps=1e-8 betas=(0.9, 0.999)
MobileViT Small	MViTs	[29]	SGD	0.1	0.0001	momentum=0.9

Table 4: Model Training information. LR: Learning rate. SGD: Stochastic gradient descent. AdamW: Adam with weight decay. RMSprop: Root mean square propagation.

378

379 A.5 Extended Few-Class Full Dataset Trained Benchmark (FC-Full) Results

380 We present the details of FC-Full results for each experiment model, including ResNet50, VGG16,
 381 ConNeXt V2 Base, Inception V3, EfficientNet V2 Medium, ShuffleNet V2, MobileNet V3 Small,
 382 ViT Base, Swin Transformer V2 Base, MobileViT Small in Fig. 3-12, respectively.

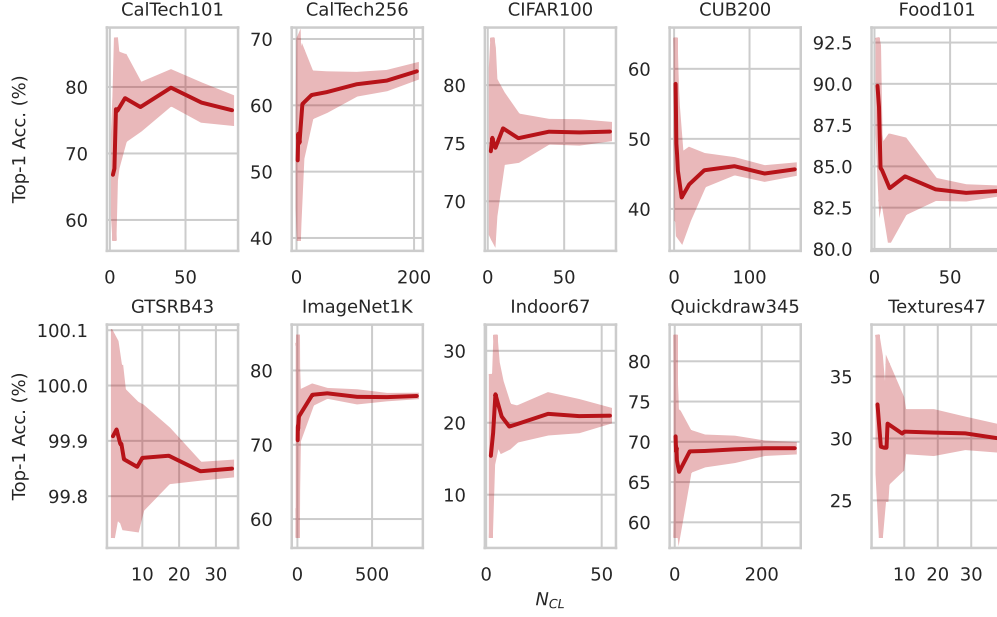


Figure 3: FC-Full Top-1 Accuracy (%) for ResNet50.

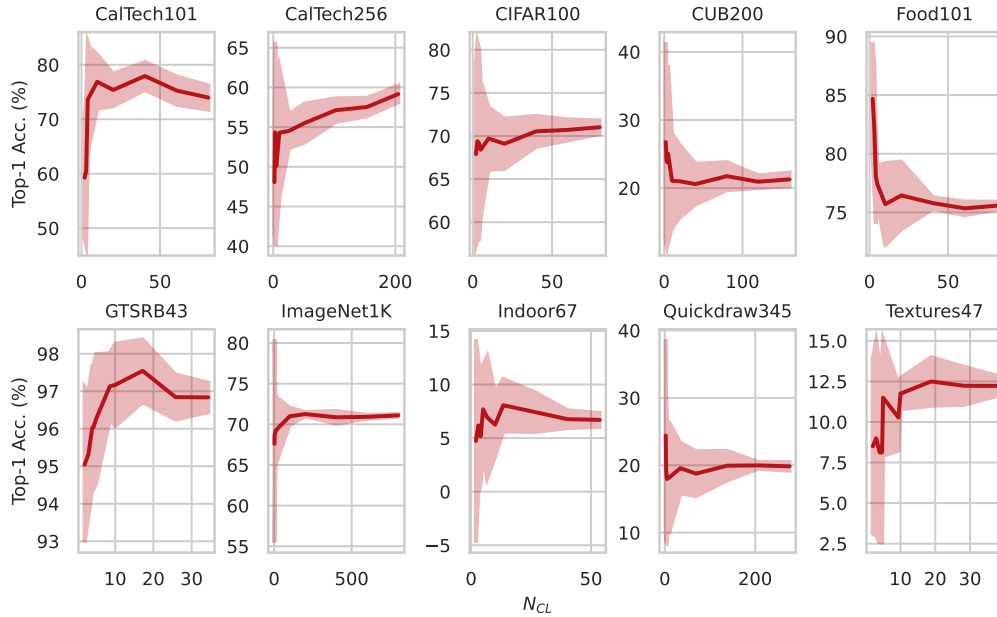


Figure 4: FC-Full Top-1 Accuracy (%) for VGG16.

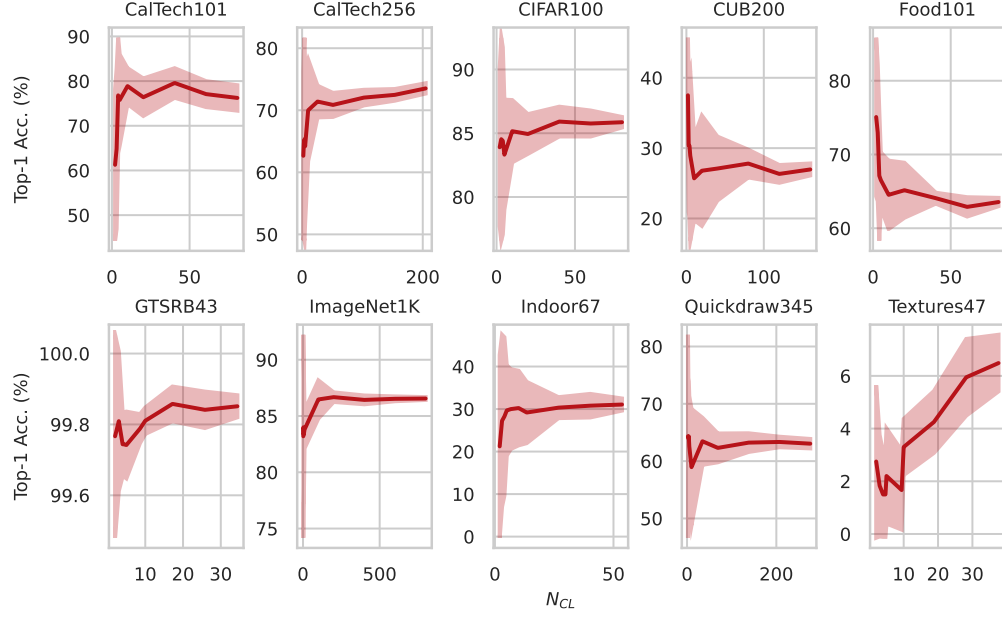


Figure 5: FC-Full Top-1 Accuracy (%) for ConNeXt V2 Base.

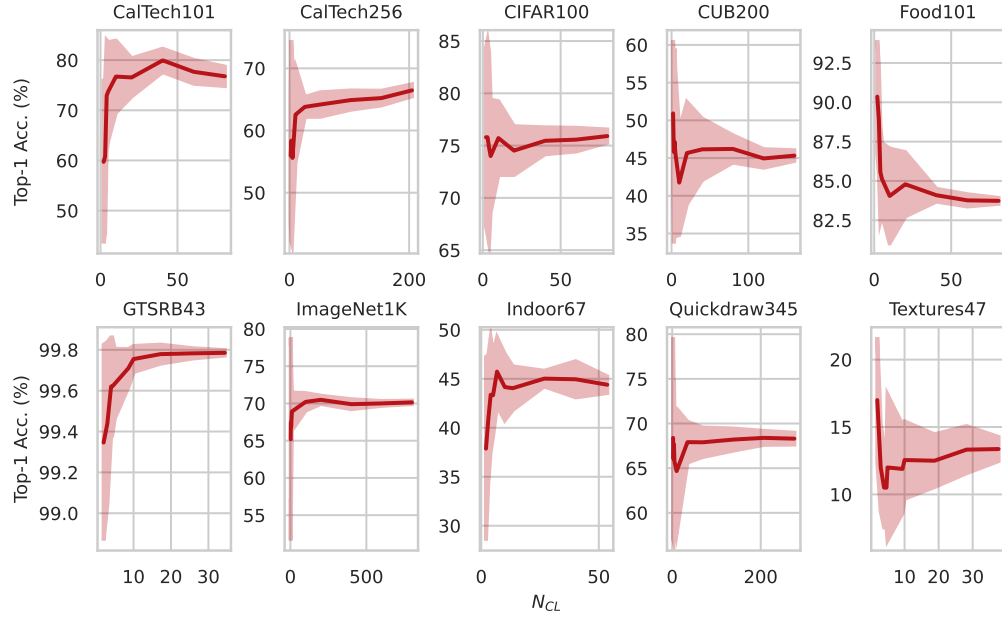


Figure 6: FC-Full Top-1 Accuracy (%) for Inception V3.

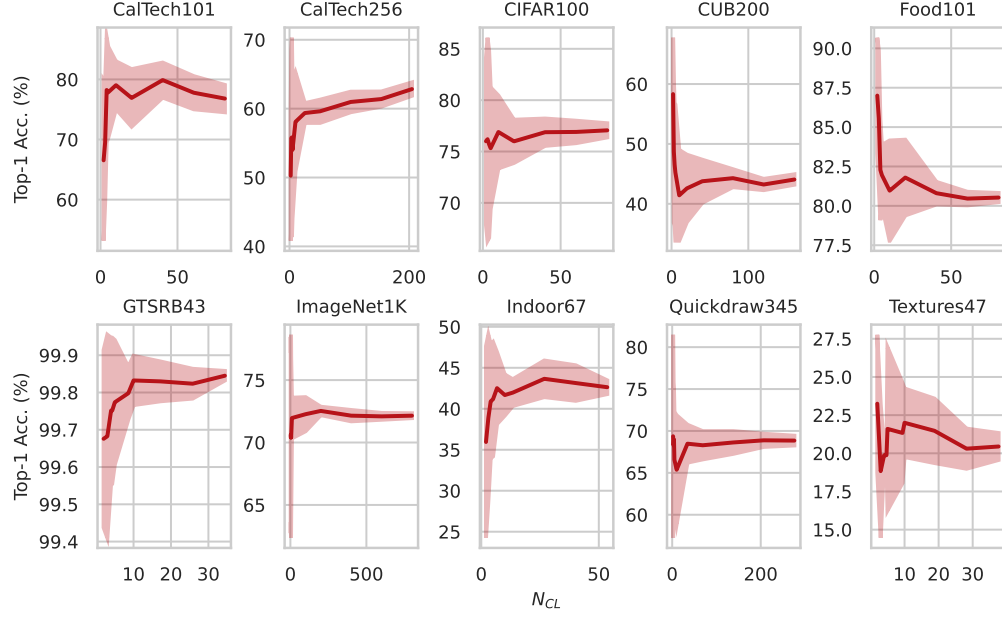


Figure 7: FC-Full Top-1 Accuracy (%) for EfficientNet V2 Medium.

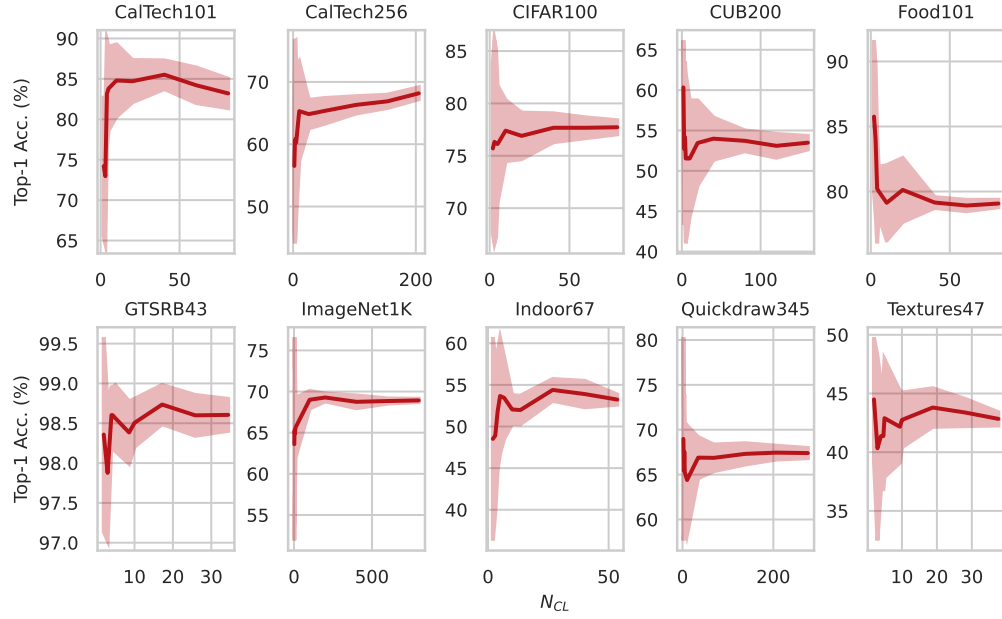


Figure 8: FC-Full Top-1 Accuracy (%) for ShuffleNet V2.

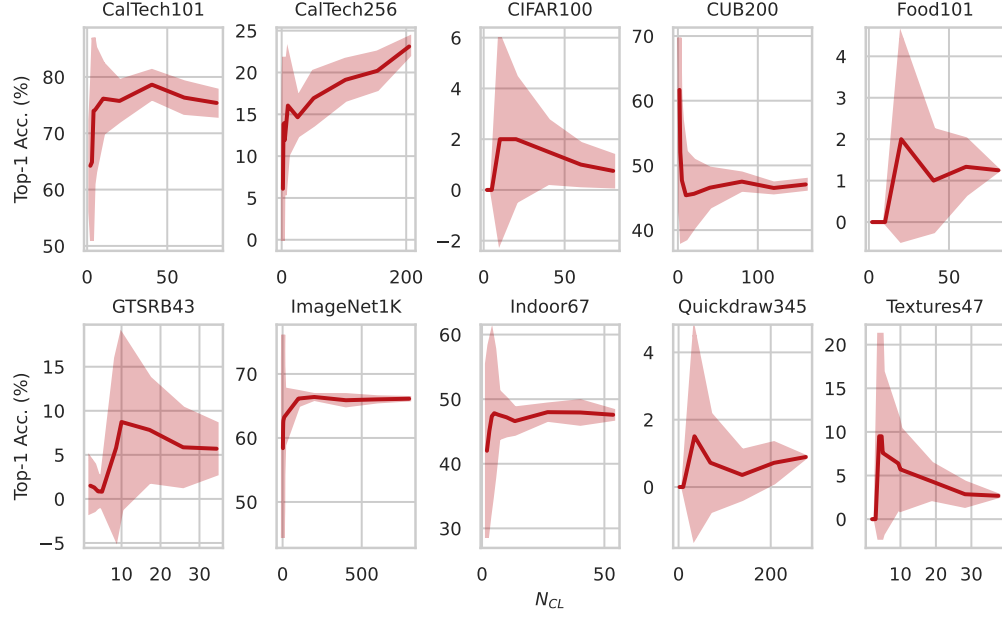


Figure 9: FC-Full Top-1 Accuracy (%) for MobileNet V3 Small.

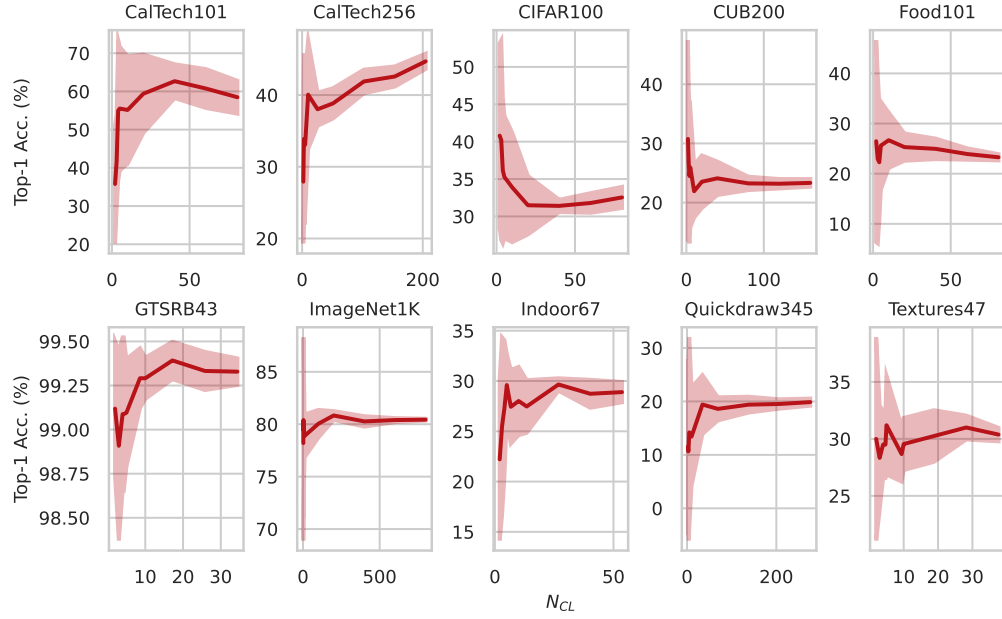


Figure 10: FC-Full Top-1 Accuracy (%) for ViT Base.

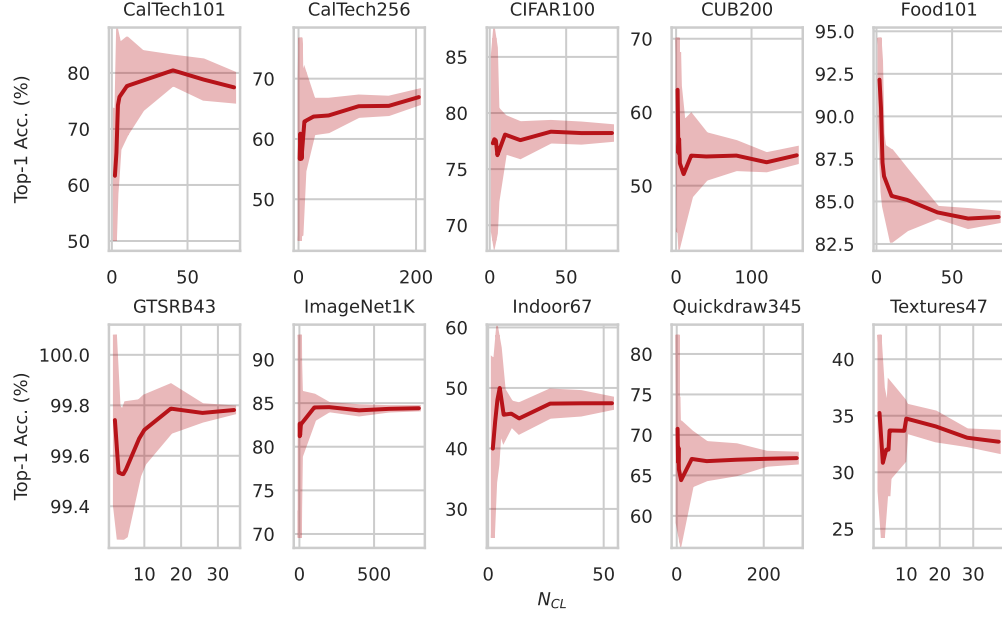


Figure 11: FC-Full Top-1 Accuracy (%) for Swin V2 Base.

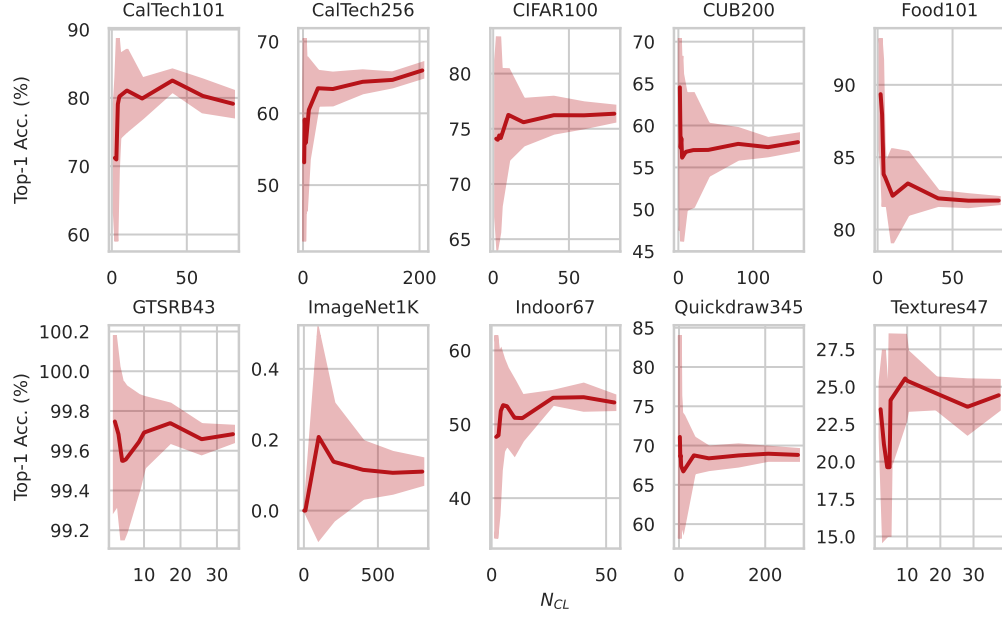


Figure 12: FC-Full Top-1 Accuracy (%) for MobileNetViT Small.

383 A.6 Extended Few-Class Subset Trained Benchmark (FC-Sub) Results

384 We present the details of FC-Sub results for each experiment model, including ResNet50, VGG16,
 385 ConNeXt V2 Base, Inception V3, EfficientNet V2 Medium, ShuffleNet V2, MobileNet V3 Small,
 386 ViT Base, Swin Transformer V2 Base, MobileViT Small in Fig. 13-22, respectively.

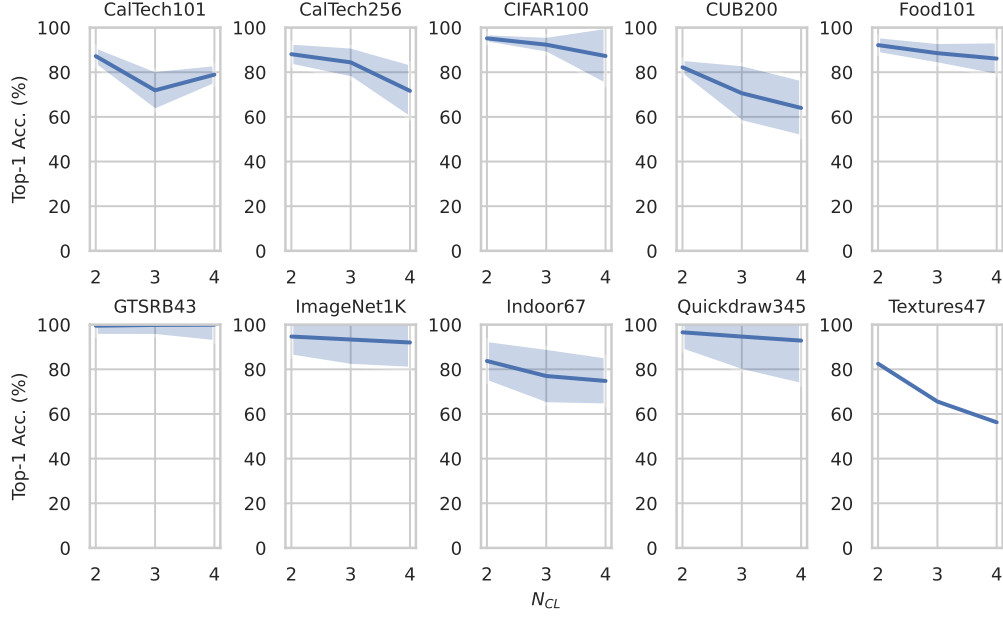


Figure 13: FC-Sub Top-1 Accuracy (%) for ResNet50.

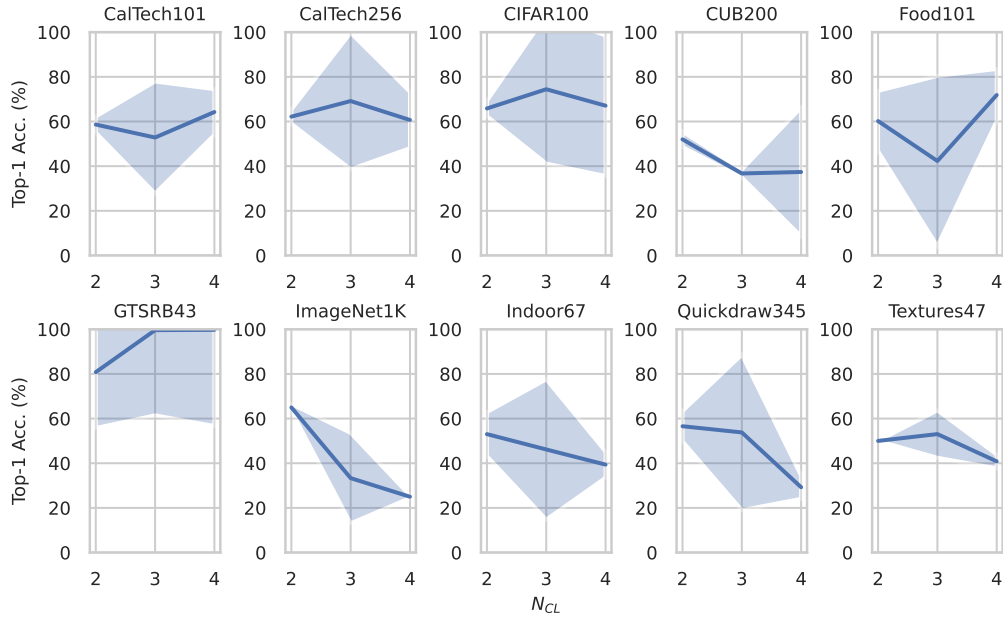


Figure 14: FC-Sub Top-1 Accuracy (%) for VGG16.

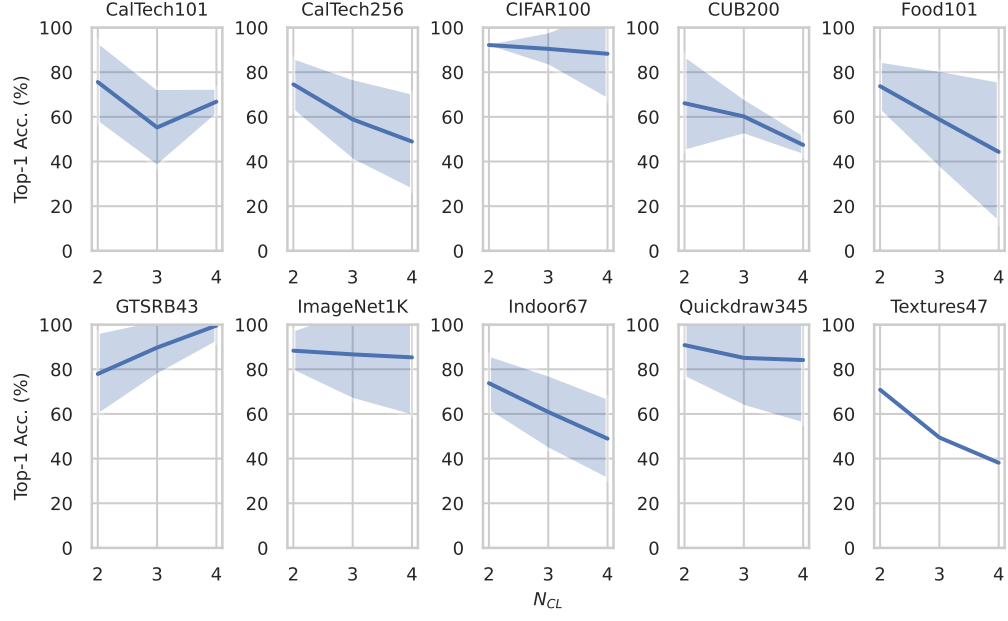


Figure 15: FC-Sub Top-1 Accuracy (%) for ConNeXt V2 Base.

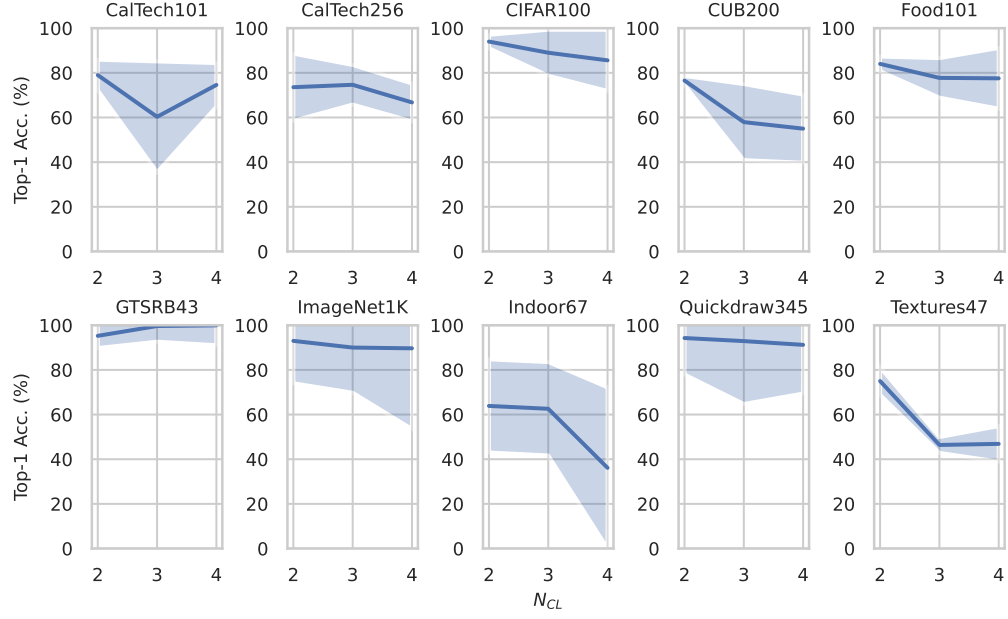


Figure 16: FC-Sub Top-1 Accuracy (%) for Inception V3.

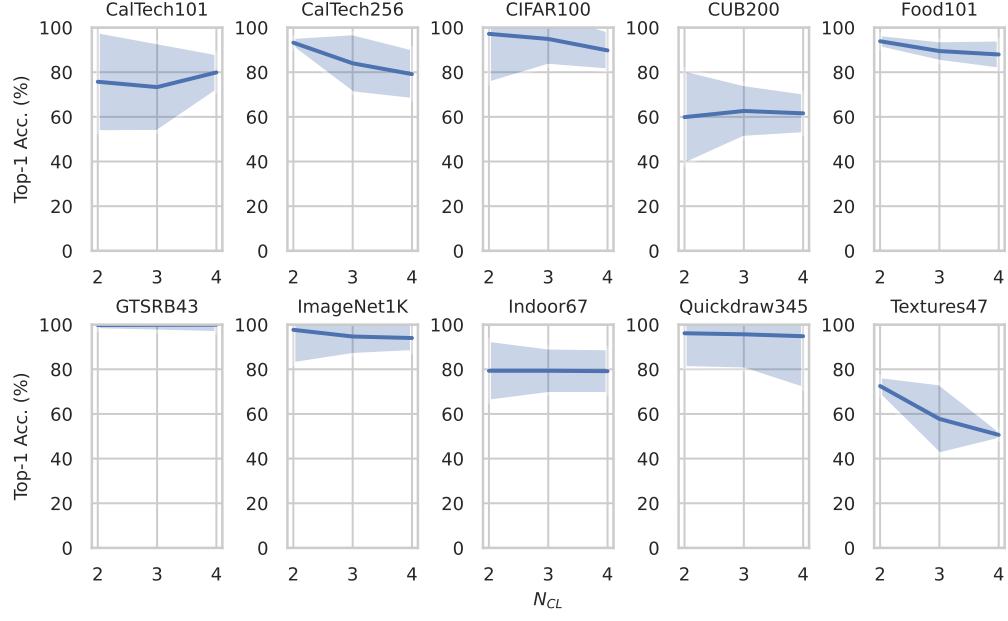


Figure 17: FC-Sub Top-1 Accuracy (%) for EfficientNet V3 Medium.

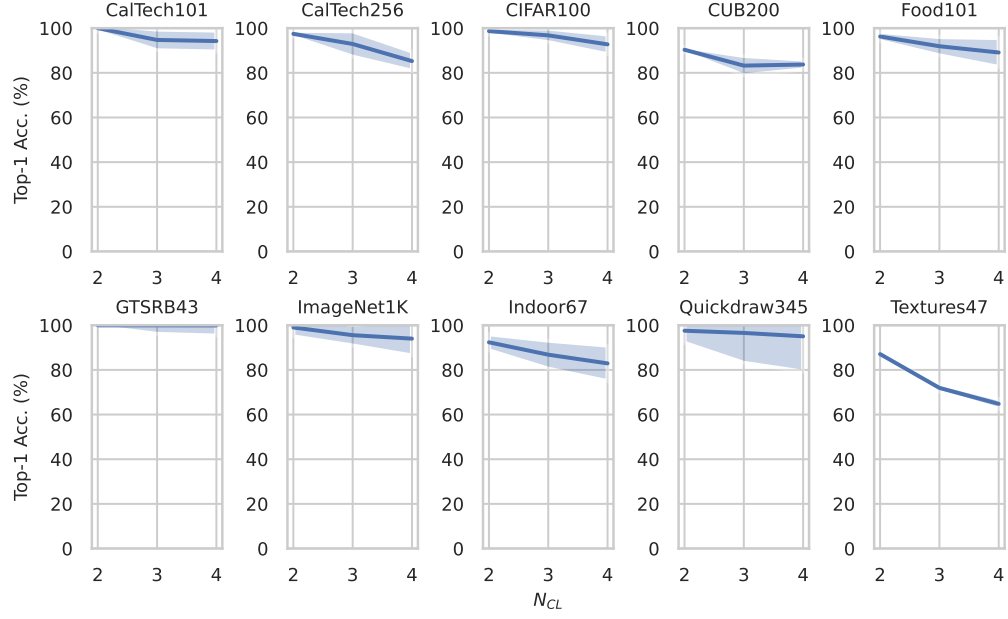


Figure 18: FC-Sub Top-1 Accuracy (%) for ShuffleNet V2.

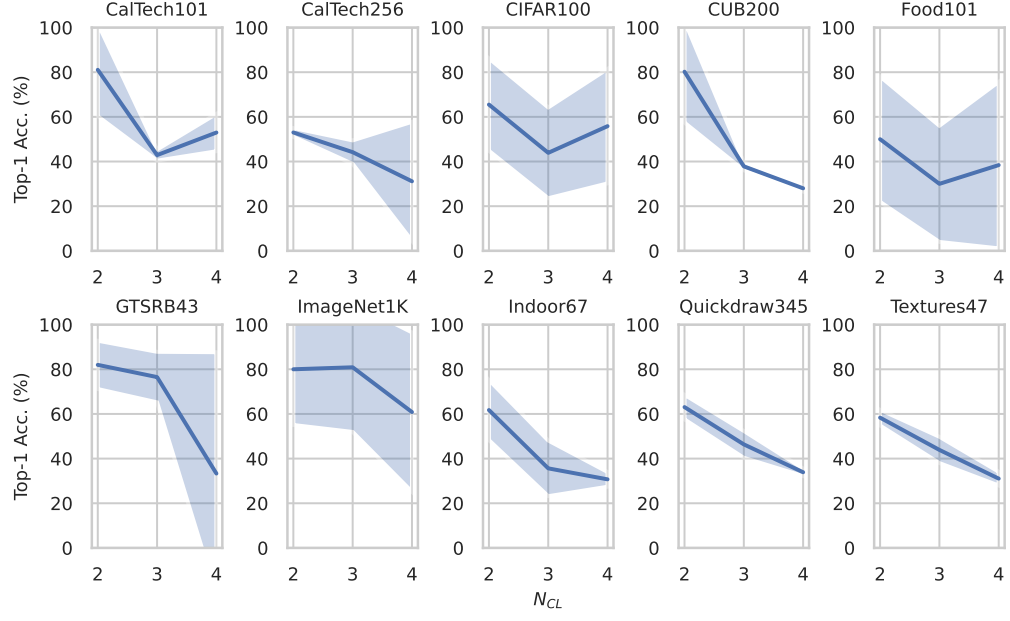


Figure 19: FC-Sub Top-1 Accuracy (%) for MobileNet V3 Small.

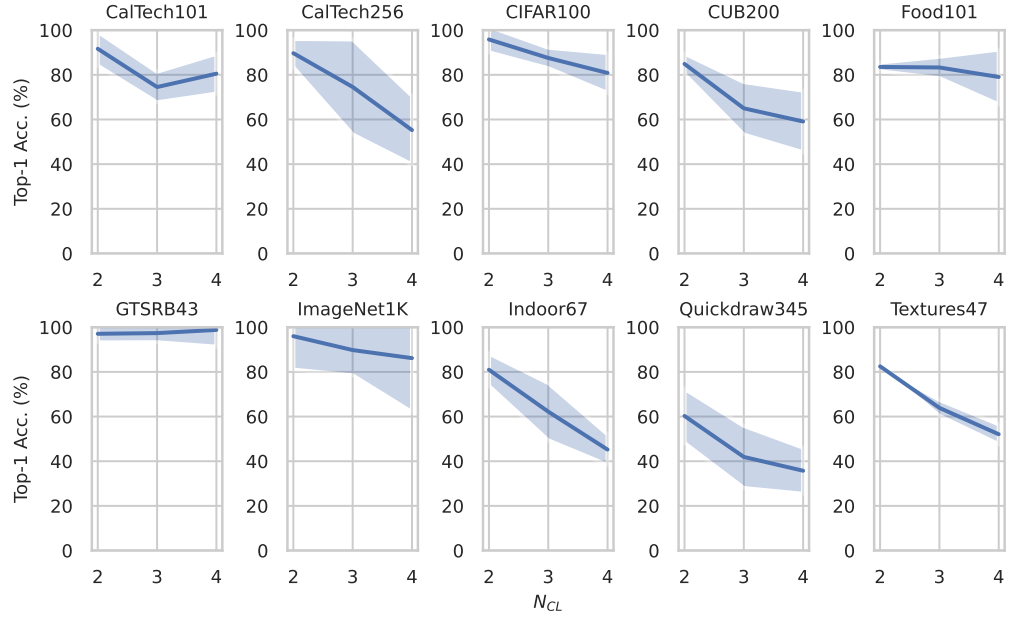


Figure 20: FC-Sub Top-1 Accuracy (%) for ViT Base.

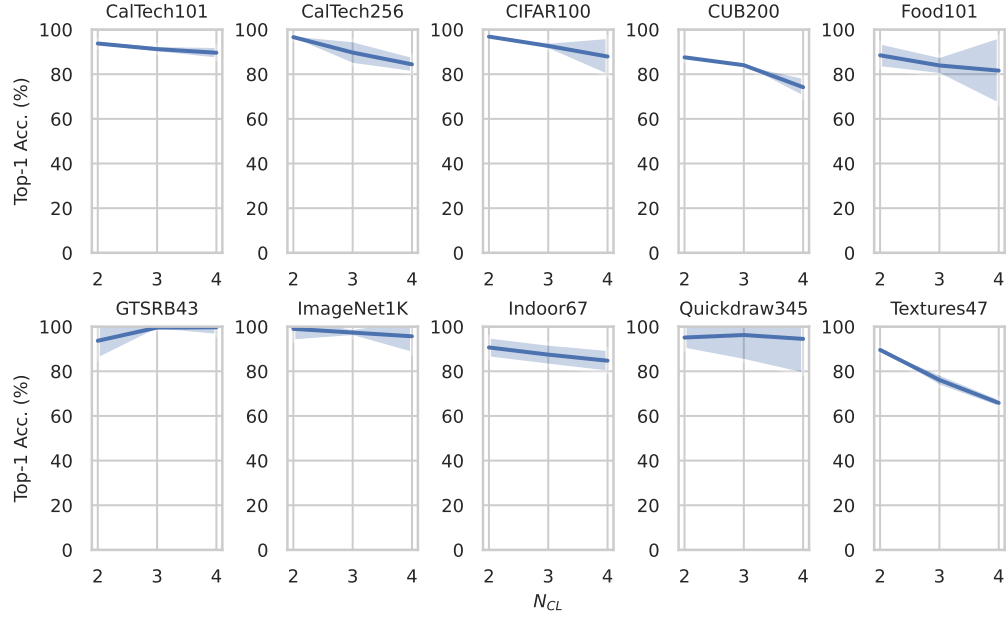


Figure 21: FC-Sub Top-1 Accuracy (%) for Swin V2.

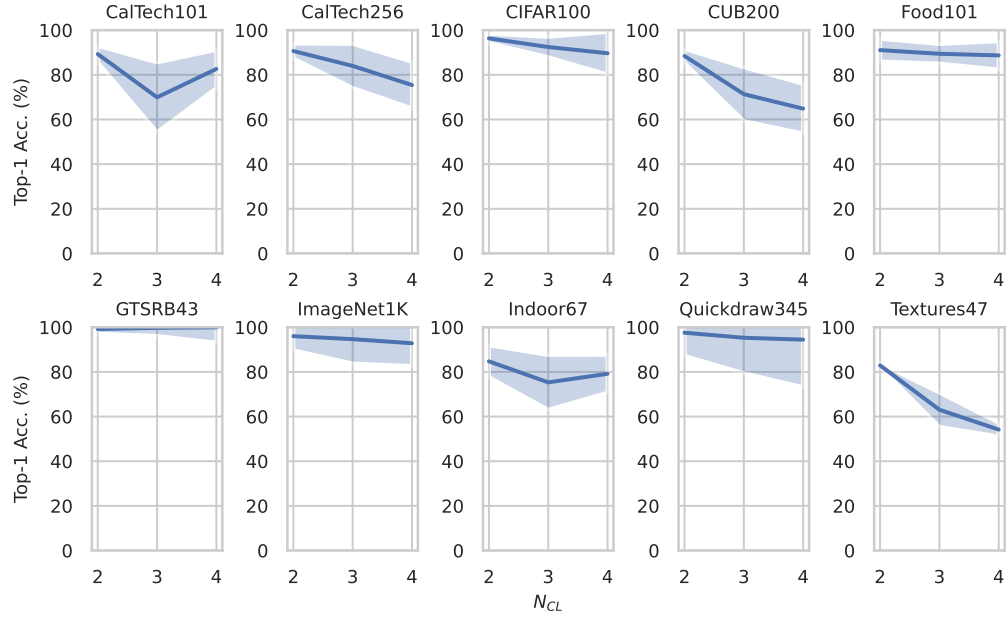


Figure 22: FC-Sub Top-1 Accuracy (%) for MobileNetViT Small.

387 A.7 Extended Few-Class Similarity Benchmark (FC-Sim) Details

388 We present the extended mathematical details of 3.6 Few-Class Similarity Benchmark (FC-Sim) in
389 the main paper.

390 The basic similarity formulation is adopted from [30]. Notations are defined as follows:

391 **Dataset D** : a set of image instances in a dataset.

392 **Class C** : a set of image instances in a class and $|C|$ is the number of instances within class C .

393 **Class Label L** : a set of class labels in a dataset and $|L|$ is the number of classes in a dataset.

394 **Feature Z_i** : visual feature of an image and i is the instance index.

395 **Class Pair $P^{(D)}$** : a set of distinct class pairs in a dataset D ; $|P^{(D)}|$ is the total number of distinct
396 class pairs.

397 **Intra-Class Image Pair $P^{(C)}$** : a set of distinct image pairs in a class C ; $|P^{(C)}|$ is the total number
398 of distinct image pairs.

399 **Inter-Class Image Pair $P^{(C_1, C_2)}$** : a set of distinct image pairs in two classes C_1, C_2 ; $|P^{(C_1, C_2)}|$ is
400 the total number of distinct image pairs. Note that this does not include same-class pairs.

401 **Intra-Class Similarity $S_\alpha^{(C)}$** : a scalar describing the similarity of images within a class by taking the
402 average of all the distinct class pairs in C :

$$S_\alpha^{(C)} = \frac{1}{|P^{(C)}|} \sum_{i, j \in C; i \neq j} \cos(\mathbf{Z}_i, \mathbf{Z}_j). \quad (1)$$

403 **Inter-Class Similarity $S_\beta^{(C_1, C_2)}$** : a scalar describing the similarity among images in two different
404 classes C_1 and C_2 :

$$S_\beta^{(C_1, C_2)} = \frac{1}{|P^{(C_1, C_2)}|} \sum_{i \in C_1, j \in C_2} \cos(\mathbf{Z}_i, \mathbf{Z}_j), \quad (2)$$

405 where C_1 and C_2 are distinct classes, i and j are the image instance indices in C_1 and C_2 , respectively.
406 $P^{(C_1, C_2)}$ is the set of distinct pairs of images between C_1 and C_2 .

407 The above equations formulate class-level similarity scores. For dataset-level, Intra-Class Similarity
408 and Inter-Class Similarity of a dataset D are defined as the mean of their similarity scores, respectively:

$$S_\alpha^{(D)} = \frac{1}{|L|} \sum_{l \in L} S_\alpha^{(C_l)} = \frac{1}{|L| \times |P^{(C_l)}|} \sum_{l \in L} \sum_{i, j \in C_l; i \neq j} \cos(\mathbf{Z}_i, \mathbf{Z}_j), \quad (3)$$

$$S_\beta^{(D)} = \frac{1}{|P^{(D)}|} \sum_{a, b \in L; a \neq b} S_\beta^{(C_a, C_b)} = \frac{1}{|P^{(D)}| \times |P^{(C_1, C_2)}|} \sum_{a, b \in L; a \neq b} \sum_{i \in C_1, j \in C_2} \cos(\mathbf{Z}_i, \mathbf{Z}_j). \quad (4)$$

409 Averaging these similarities can provide a summary of score in class or dataset levels by a single
410 scalar. However, this simplicity neglects other cluster-related information that can better reveal the
411 underlying dataset difficulty property of a dataset. In particular, the **(1) tightness of a class cluster**
412 and **(2) distance to other classes** of class clusters, are features that characterize the inherent class
413 difficulty, but are not captured by S_α or S_β alone.

414 To compensate the aforementioned drawback, we adopt the Silhouette Score (SS) (also called
415 Silhouette Coefficient in the literature) [31, 32]:

$$SS(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}, \quad (5)$$

416 where $SS(i)$ is the Silhouette Score of the data point i , $a(i)$ is the average dissimilarity between i and
417 other instances in the same class, and $b(i)$ is the average dissimilarity between i and other data points

418 in the closest different class. Intuitively, this metric summarizes the quality of clusters jointly by the
 419 degree of instances of the same class and distinct clusters, normalized by the longest distance of $a(i)$
 420 and $b(i)$. By this definition, we can see that $-1 \leq SS(i) \leq 1$ where -1 indicates a dataset is poorly
 421 clustered (data points with different classes are scattered around) while 1 represents a well-clustered
 422 dataset.

423 Euclidean Distance is commonly used to measure two data points' differences; in contrast, we incor-
 424 porate the inverse of similarity (dissimilarity) as data point's difference into the existing Silhouette
 425 Score. Observe that the above Intra-Class Similarity $S_\alpha^{(C)}$ already represents the tightness of the class
 426 (C), therefore $a(i)$ can be replaced with the inverse of Intra-Class Similarity $a(i) = -S_\alpha(i)$. For the
 427 second term $b(i)$, we adopt the previously defined Inter-Class Similarity $S_\beta^{(C_1, C_2)}$ and introduce a
 428 new similarity score as follows:

429 **Nearest Inter-Class Similarity** $S'_\beta^{(C)}$: a scalar describing the similarity among instances between
 430 class C and the closest class of each instance in C :

$$S'_\beta^{(C)} = \frac{1}{|P(C, \hat{C})|} \sum_{i \in C, j \in \hat{C}} \cos(\mathbf{Z}_i, \mathbf{Z}_j), \quad (6)$$

431 where \hat{C} is the set of the nearest class to C ($\hat{C} \neq C$).

432 Consequently, the dataset-level Nearest Inter-Class Similarity $S'^{(D)}_\beta$ is expressed as:

$$S'^{(D)}_\beta = \frac{1}{|L|} \sum_{l \in L} S'^{(C_l, \hat{C}_l)}_\beta = \frac{1}{|L| \times |P(C_l, \hat{C}_l)|} \sum_{l \in L} \sum_{i \in C_l, j \in \hat{C}_l} \cos(\mathbf{Z}_i, \mathbf{Z}_j). \quad (7)$$

433 The second term of $SS(i)$ can be written as $b(i) = -S'_\beta(i)$.

434 Replacing $a(i)$ and $b(i)$ from equation 5 with these similarity terms, we introduce our novel similarity
 435 metric:

436 **Similarity-Based Silhouette Score** $SimSS$:

$$SimSS(i) = \frac{S_\alpha(i) - S'_\beta(i)}{\max(S_\alpha(i), S'_\beta(i))}, \quad \text{for instance } i \quad (8)$$

437

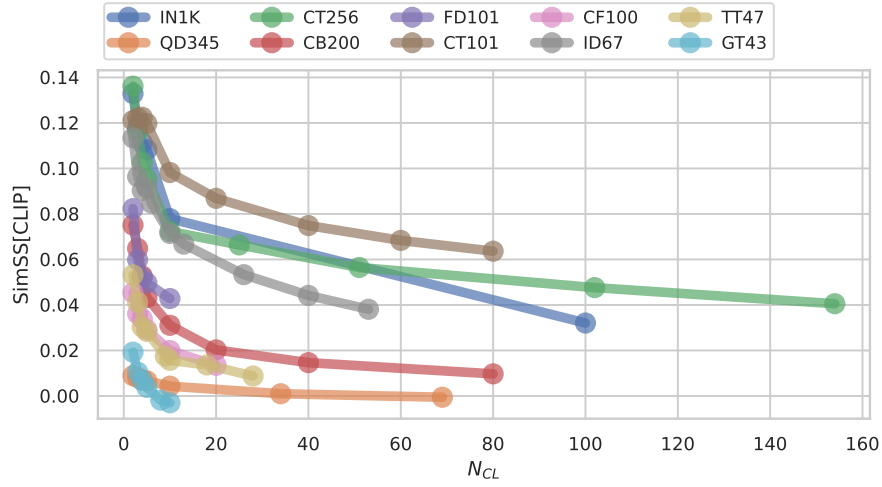
$$SimSS^{(C)} = \frac{1}{|C|} \sum_{i \in C} SimSS(i) = \frac{1}{|C|} \sum_{i \in C} \frac{S_\alpha(i) - S'_\beta(i)}{\max(S_\alpha(i), S'_\beta(i))}, \quad \text{for class } C \quad (9)$$

$$\begin{aligned} SimSS^{(D)} &= \frac{1}{|L|} \sum_{l \in L} SimSS^{(C_l)} = \frac{1}{|L| \times |C_l|} \sum_{l \in L, i \in C_l} SimSS(i) \\ &= \frac{1}{|L| \times |C_l|} \sum_{i \in C_l} \frac{S_\alpha(i) - S'_\beta(i)}{\max(S_\alpha(i), S'_\beta(i))}, \quad \text{for dataset } D. \end{aligned} \quad (10)$$

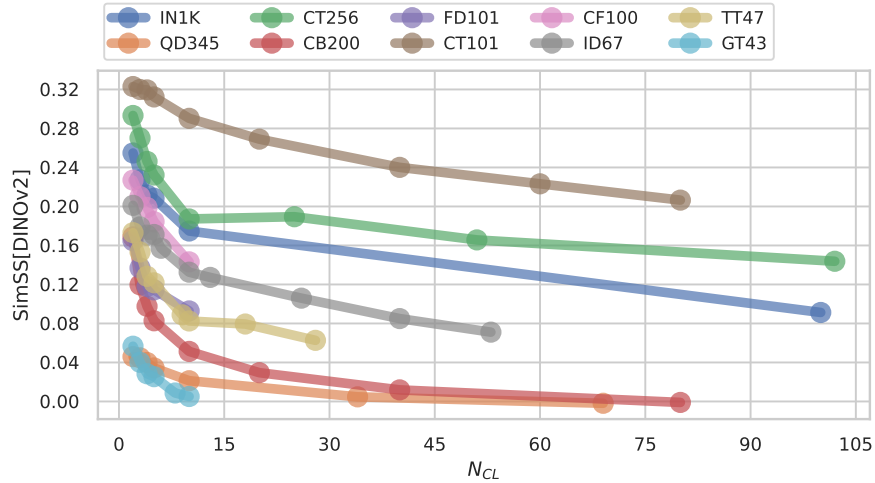
438 A.8 Extended Few-Class Similarity Benchmark (FC-Sim) Results

439 We present the relationship of similarity scores using our proposed SimSS and number of classes
 440 N_{CL} in ten datasets. CLIP and DINOv2 are used as similarity base functions of SimSS shown in Fig.
 441 23 (a) and (b), respectively.

442 Overall, a key observation is that the general trend among all ten datasets unveils the inverse
 443 relationship between similarity and the number of classes. Specifically, image similarities, which act
 444 as proxy of inverse subset difficulty score increases as the number of classes N_{CL} decreases. This
 445 reveals that similarity plays a more important role in the *Few-Class Regime* than for datasets with
 446 more classes. Therefore, for real applications with few classes, simply downscaling a model blindly
 447 without considering class similarity may yield a model selection with sub-optimal efficiency. We
 448 propose, therefore, that image similarity must be taken into consideration for existing scaling laws
 449 [33, 34, 35]. To that end, *Few-Class Arenas* developed to facilitate future research in this direction.



(a) SimSS using CLIP as similarity base function vs N_{CL} curve.



(b) SimSS using DINOv2 as similarity base function vs N_{CL} curve.

Figure 23: Relation of SimSS[CLIP,DINOv2] and N_{CL} .

450 Note that both CLIP and DINOv2 are trained on images from the Internet similar to ImageNet.
 451 Therefore to what extent they can capture image similarity in different types is an open research

question. Examples include drawings without textures in the Quickdraw dataset (QD345), textures without shapes in the Describable Textures Dataset (TT47), etc. We mentioned this limitation also in the main manuscript.

Effect of ResNet Scales on Similarity. We present the details of FCA-Sim results of the ResNet family in different scales in the *Few-Class Regime* of ImageNet1K, specifically ($N_{CL} \in \{2, 3, 4, 5, 10, 100\}$) shown in Fig. 24. In particular, we analyze the relationship between each full and sub-Model’s Top-1 accuracy and SimSS by Pearson correlation coefficient (PCC) denoted as r in the plots. The ResNet family scales from ResNet18 to ResNet152. We experiment both CLIP (red in the first two rows) and DINOv2 (blue in the last two rows) as similarity base functions.

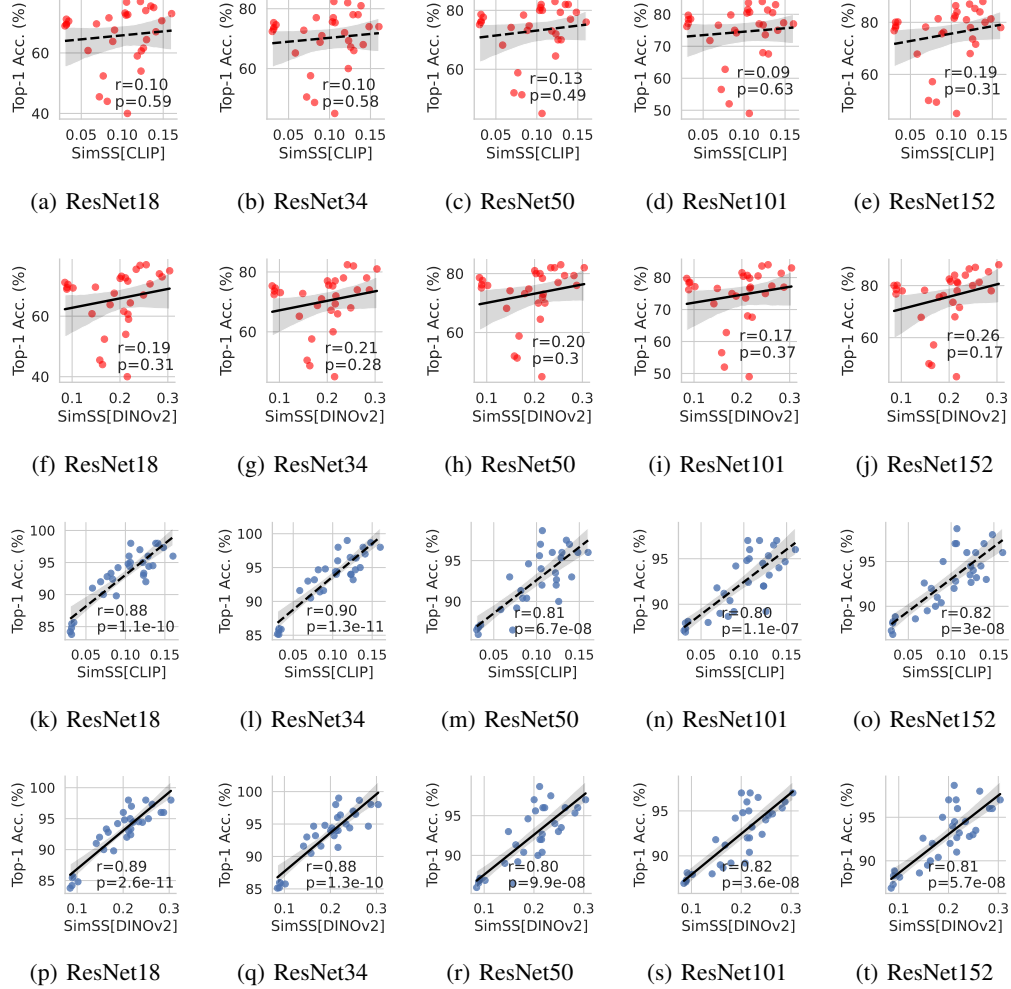


Figure 24: Top-1 Accuracy of different ResNet scales vs SimSS on ImageNet1K. $N_{CL} \in \{2, 3, 4, 5, 10, 100\}$.

460

In general, ResNet presents high correlation ($r \geq 0.80$) between sub-models’ performances and SimSS (blue in the last two rows), compared to full models’ performances and SimSS ($r \leq 0.26$, red in the first two rows). This high correlation indicates that SimSS can be used as a reliable tool to estimate upper bound accuracies of ResNet sub-models. Comparing CLIP (r in the 1st row) with DINOv2 (r in the 2nd row) as similarity base functions, observe that PCC is slightly higher for DINOv2 on full models than CLIP, while these differences are subtle for sub-models (r in the 3rd row vs 4th row). Regarding sub-models of ResNet in different scales, the two smallest models’ accuracies (ResNet18 and ResNet34) have higher correlation with SimSS ($r \geq 0.88$), compared to larger models

with (ResNet50, ResNet101 and ResNet152) $r \geq 0.80$. We opens a new direction of novel scaling law considering image similarity for efficient models in *Few-Class Regime*.

Transformer vs Similarity. We present the details of Vision Transformer (ViT) [11] Base performance and SimSS using CLIP and DINOv2 in the subsets of $N_{CL} \in \{2, 3, 4, 5, 10, 100\}$ in ImageNet1K, shown in Fig. 25. An overview of the results indicates low correlation between Top-1 accuracy and SimSS ($r \leq 0.20$). Our tool *Few-Class Arena* helps discover the new challenge of using the vanilla ViT in the *Few-Class Regime*, as ViTs are originally designed to scale up [35] when large training datasets in the order of billion samples are available. *Few-Class Arena* also identifies the importance of scaling down considering similarity.

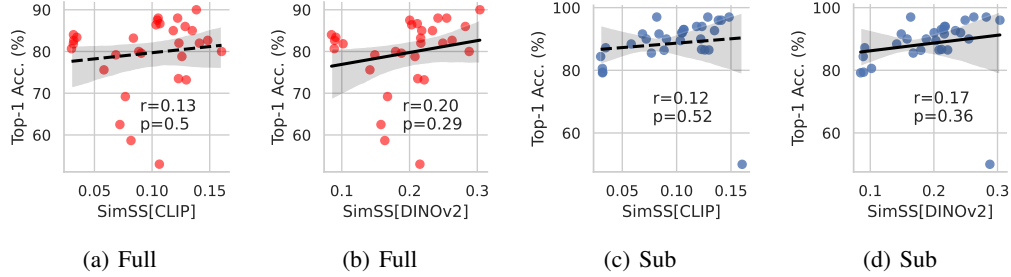


Figure 25: Top-1 Accuracy of ViT Base vs SimSS on ImageNet1K. $N_{CL} \in \{2, 3, 4, 5, 10, 100\}$.

A.9 Experiments Compute Resources

Experiments are conducted in two internal clusters with the following specifications: (1) 8 NVIDIA RTX A5000 24GB, AMD EPYC 7513 32-Core Processor 882GB RAM and (2) 8 NVIDIA TITAN Xp 12GB, Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz, 126GB RAM. When GPUs in two clusters are fully utilized, training 10 models in 9 datasets takes 2 weeks; obtaining one experiment result of FC-Full usually takes less than 1 minute since it only involves inference without training; obtaining one experiment result of FC-Sub takes around 2 days on average depending on the size of subset and model, which includes both training and testing; computing the SimSS in the *Few-Class Regime* for ten datasets takes around three weeks.

References

- [1] Contributors, M. Openmmlab’s pre-training toolbox and benchmark. <https://github.com/open-mmlab/mmpretrain>, 2023.
- [2] Peng, Z., W. Huang, S. Gu, et al. Conformer: Local features coupling global representations for visual recognition. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 357–366, 2021.
- [3] Liu, Z., H. Mao, C. Wu, et al. A convnet for the 2020s. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976, 2022.
- [4] Li, Y., G. Yuan, Y. Wen, et al. Efficientformer: Vision transformers at mobilenet speed. *ArXiv*, abs/2206.01191, 2022.
- [5] Tan, M., Q. V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *ArXiv*, abs/1905.11946, 2019.
- [6] Ding, X., X. Zhang, N. Ma, et al. Repvgg: Making vgg-style convnets great again. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13728–13737, 2021.
- [7] He, K., X. Zhang, S. Ren, et al. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778. 2016.

- [8] Zhang, X., X. Zhou, M. Lin, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6848–6856, 2017.
- [9] Liu, Z., Y. Lin, Y. Cao, et al. Swin transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021.
- [10] Simonyan, K., A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.
- [11] Dosovitskiy, A., L. Beyer, A. Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [12] Li, F.-F., M. Andreeto, M. Ranzato, et al. Caltech 101, 2022.
- [13] Griffin, G., A. Holub, P. Perona. Caltech 256, 2022.
- [14] Krizhevsky, A., G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [15] Wah, C., S. Branson, P. Welinder, et al. *The Caltech-UCSD Birds-200-2011 Dataset*. 2011.
- [16] Bossard, L., M. Guillaumin, L. Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*. 2014.
- [17] Stallkamp, J., M. Schlipsing, J. Salmen, et al. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, (0):–, 2012.
- [18] Deng, J., W. Dong, R. Socher, et al. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [19] Quattoni, A., A. Torralba. Recognizing indoor scenes. In *2009 IEEE conference on computer vision and pattern recognition*, pages 413–420. IEEE, 2009.
- [20] Ha, D., D. Eck. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*, 2017.
- [21] Cimpoi, M., S. Maji, I. Kokkinos, et al. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2014.
- [22] Simonyan, K., A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [23] Woo, S., S. Debnath, R. Hu, et al. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16133–16142. 2023.
- [24] Szegedy, C., V. Vanhoucke, S. Ioffe, et al. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826. 2016.
- [25] Tan, M., Q. L. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.
- [26] Ma, N., X. Zhang, H.-T. Zheng, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131. 2018.
- [27] Howard, A., M. Sandler, G. Chu, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324. 2019.
- [28] Liu, Z., H. Hu, Y. Lin, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019. 2022.
- [29] Mehta, S., M. Rastegari. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. arxiv 2021. *arXiv preprint arXiv:2110.02178*.

- 550 [30] Cao, B. B., L. O’Gorman, M. Coss, et al. Data-side efficiencies for lightweight convolutional
551 neural networks. *arXiv preprint arXiv:2308.13057*, 2023.
- 552 [31] Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster
553 analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.
- 554 [32] Shahapure, K. R., C. Nicholas. Cluster quality analysis using silhouette score. In *2020 IEEE*
555 *7th international conference on data science and advanced analytics (DSAA)*, pages 747–748.
556 IEEE, 2020.
- 557 [33] Kaplan, J., S. McCandlish, T. Henighan, et al. Scaling laws for neural language models. *arXiv*
558 *preprint arXiv:2001.08361*, 2020.
- 559 [34] Rae, J. W., S. Borgeaud, T. Cai, et al. Scaling language models: Methods, analysis & insights
560 from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.
- 561 [35] Zhai, X., A. Kolesnikov, N. Houlsby, et al. Scaling vision transformers. In *Proceedings of the*
562 *IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113. 2022.