Filter, Recognize, Organize: A Four-Stage Pipeline for Fine-Grained Recognition on a Web-Scraped Archive of Cheese Images

***Abstract*—Web-scraped image search enables rapid construction of large-scale fine-grained datasets but introduces systematic noise in both visual content and label structure. We study these issues using a dataset of 165,769 images spanning 1,824 cheese-named categories. We propose a four-stage pipeline consisting of (i) a CLIP-based filter to remove packaging-dominant images, (ii) fine-grained recognition over 1,238 filtered categories using ConvNeXt, (iii) evaluation under an external cheese taxonomy to assess hierarchical consistency, and (iv) name-based canonicalization to merge near-duplicate categories. Across controlled experiments using a fixed train/validation/test split, we compare the effects of backbone scaling, loss reweighting, and label-space refinement. We find that label-space refinement yields larger improvements than moderate backbone scaling, merging near-duplicate categories improves top-1 accuracy by 5.2 percentage points, while increasing model capacity from ConvNeXt-Tiny to ConvNeXt-Small under an identical 25-epoch training budget yields a 0.2-point gain. Class-balanced loss improves minority-class performance but reduces overall accuracy under the given test distribution. These results indicate that, under the evaluated model scales and dataset conditions, label structure is a major limiting factor in this regime, compared with moderate increases in model capacity.**

***Index Terms*—fine-grained visual recognition, food recognition, CLIP, transfer learning, label noise, long-tailed recognition, dataset curation**

I. Introduction

Web image search is frequently used to construct large-scale labeled datasets by issuing category names as queries. While this approach is inexpensive and scalable, the resulting datasets inherit structural noise. In fine-grained domains, such noise is particularly harmful because inter-class variation is small and label semantics are often inconsistent. For a query such as *"brie cheese"*, a large fraction of returned images depict supermarket packaging dominated by brand logos and text rather than the food itself, and category names such as *"cheddar"*, *"mature cheddar"*, and *"orkney extra mature cheddar"* are treated as separate classes despite being highly similar in visual appearance at the resolution of web images.

We study a public web-scraped cheese image archive containing 165,769 images organized into 1,824 cheese-named categories. The dataset exhibits three main issues: (i) off-target images such as packaging and branding-heavy photographs, (ii) long-tailed class distributions, and (iii) inconsistent label granularity, including near-duplicate category names that refer to products difficult to distinguish at the resolution of web images.

We analyze how performance changes relate to data curation, and how the label space varies with the evaluation protocol. The train/validation/test split is held fixed, allowing the gains attributable to representation learning to be separated from those attributable to dataset design.

Our contributions are:

- A four-stage pipeline for cleaning and reorganizing a noisy web-scraped fine-grained dataset, comprising packaging filtering, type recognition, family-level organization, and near-duplicate label consolidation.
- An empirical evaluation on the publicly available Cheese Pics dataset, supporting reproducibility and future comparison.
- A controlled comparison of backbone scaling, label-space refinement, and loss reweighting under a fixed split.
- Evidence that, for this dataset and the model scales examined, label-space canonicalization yields larger gains than moderate increases in backbone capacity, improving top-1 accuracy from scaling ConvNeXt-Tiny to ConvNeXt-Small at an identical training budget.

II. Related Work

Vision-language pretraining. Vision-language pretraining methods such as CLIP [1], learn aligned image–text representations that transfer well to downstream tasks. We employ a frozen ViT-L/14 encoder [2] via OpenCLIP [3] for feature extraction and filtering. Zero-shot classification by comparing image features to textual class prompts is attractive. But, as our results indicate, it is limited to rare category names.

Fine-grained visual categorization (FGVC). FGVC benchmarks such as CUB-200 [4], iNaturalist [5], and Food-101 [6] focus on curated label spaces with relatively clean supervision, and strong results typically come from fine-tuning a high-capacity backbone such as ConvNeXt [7]. In contrast, our setting emphasizes web-derived label noise and inconsistent category granularity.

Learning from web supervision and label noise. WebVision [8] studies learning from noisy web labels and highlights the importance of data quality. Our study complements this work by explicitly quantifying the effect of label-space restructuring.

Long-tailed recognition. Long-tailed recognition methods, including class-balanced loss [9], improve minority-class accuracy but may reduce overall accuracy under naturally imbalanced test distributions.

TABLE I
ARCHIVE AND LABEL-SET STATISTICS.

Quantity	Value
Folders (cheese names)	1,824
Total readable images	165,769
Hand labels for Stage 1 (yes / no)	105 / 103
Images kept after packaging filter	62,174 (37.5%)
Cheeses with ≥ 20 clean images (Stage 2 set)	1,238
Images in Stage 2 set (capped at 40/type)	43,489
Mean clean images per cheese	34

Dataset curation and de-duplication. Dataset curation techniques, including de-duplication and label cleaning, are known to improve performance. Still, their effect relative to model scaling is less often isolated in controlled comparisons, particularly in fine-grained food domains. Our near-duplicate problem is at the *label* level. Distinct folders denote the same physical product, so the categories are not reliably separable from image content. We address it with a name-based canonicalization grounded in an external catalog and quantify its effect against the alternative of scaling the model.

III. THE ARCHIVE

The dataset is derived from the publicly available Cheese Pics collection on Kaggle [10]. It contains approximately 166,000 images collected via web search and organized into folders corresponding to cheese names. Because the dataset is publicly available, all experiments in this paper can be reproduced using the same source data. Our contribution is not the collection of new images but the systematic filtering, reorganization, and evaluation of this existing web-scraped archive.

After removing 8 unreadable files, we obtain 165,769 images across 1,824 folders. Folder sizes are highly uneven, with a median near 90 images. Two dominant issues are observed.

Packaging dominance. A substantial fraction of images depict retail packaging rather than the underlying cheese, introducing visual confounds dominated by text and branding that are nearly identical across unrelated categories.

Label redundancy and inconsistency. Many category names differ only by modifiers such as maturity or origin descriptors (*"mature cheddar"*, *"orkney extra mature cheddar"*) while corresponding to products that are highly similar in appearance at the resolution of web images. After Stage 1 filtering, 1,798 categories retain at least one clean image, 1,541 at least ten, and 1,238 at least twenty (median 31, mean 34). We retain the 1,238 categories with at least 20 clean images as the recognition set (Fig. 1). This subset remains long-tailed and contains systematic near-duplicates, which motivates the label-space consolidation in Stages 3 and 4.

For the family hierarchy (Stage 3), we align cheese names to a public catalog derived from *cheese.com* [11], which provides a curated texture-based `type` field (e.g., *soft*, *semi-hard*, *hard*, *blue-veined*).

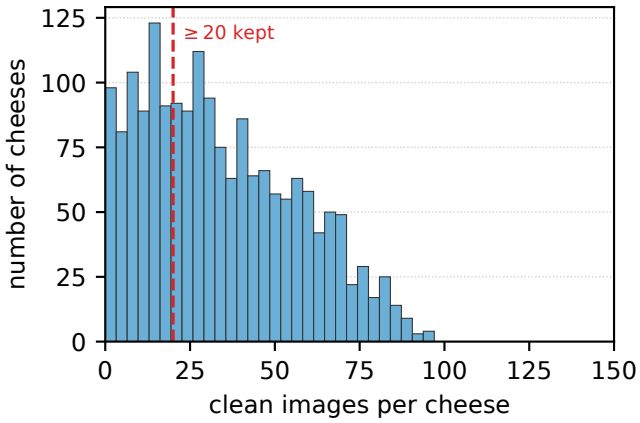

Fig. 1. Distribution of clean (post-filter) images per cheese. The corpus is long-tailed. The dashed line marks the $\geq$20-image threshold that defines the 1,238-class recognition set (counts above 150 are clipped into the last bin).

IV. METHOD

Figure 2 summarizes the four stages. All CLIP features are 768-dimensional, L_2-normalized embeddings from ViT-L/14.

A. Stage 1: Packaging Filter

The filter is formulated as a binary classification task, whether an image shows real, unpackaged cheese ("good") or commercial packaging ("not_good"). We hand-labeled 208 images (105 good, 103 packaging) and extracted CLIP features for them. A logistic regression with standardized features and balanced class weights is fit, with the inverse-regularization C chosen by 5-fold stratified cross-validation ($C = 0.01$). The probe is then run over all 165,769 images. It labels 62,174 (37.5%) as good. Despite limited supervision, the model achieves strong separation in feature space, reflecting the near-linear separability of packaging and food regions in CLIP feature space.

B. Stage 2: Fine-Grained Type Recognition

On the 1,238-cheese clean set, we compare three approaches of increasing cost.

Zero-shot CLIP. Each image is matched to the textual prompt *"a photo of* {*name*}*"* for every cheese. The argmax is the prediction.

Linear probe. A logistic-regression head is trained on frozen CLIP features, evaluated by 3-fold stratified cross-validation.

Fine-tuned backbone. We fine-tune an ImageNet-pretrained ConvNeXt [7] end-to-end. Training uses AdamW [12] (lr $= 10^{-4}$, weight decay 0.05), cosine schedule, label smoothing 0.1, mixed precision, RandAugment [13], and a class-balanced 224×224 random-resized-crop pipeline. We use a stratified 70/15/15 train/val/test split *per cheese*, fixed once and reused across all experiments, so results are directly comparable. We report ConvNeXt-Tiny and the larger ConvNeXt-Small, both trained for 25 epochs, on the identical split.

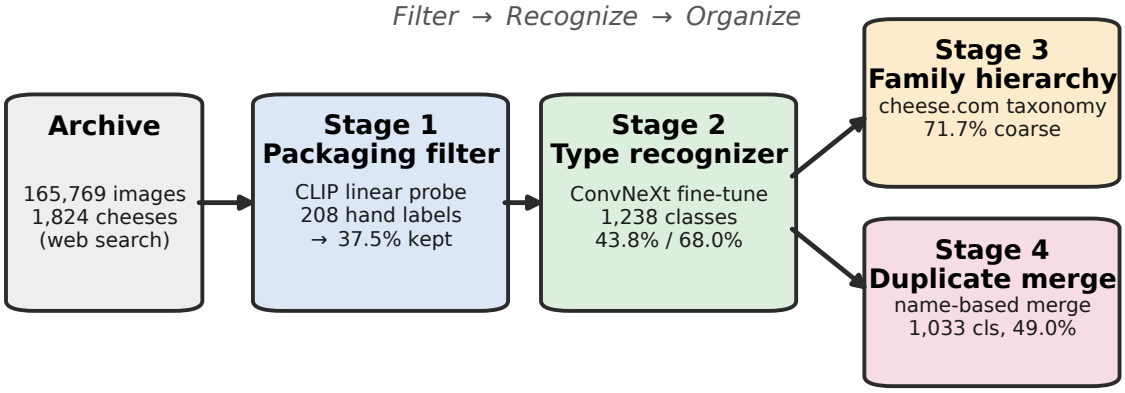

Fig. 2. The four-stage pipeline. Stage 1 keeps food images; Stage 2 recognizes the cheese type; Stages 3 and 4 reorganize the label space (family roll-up and near-duplicate merge) to recover accuracy that the raw label space limits. Annotations show the headline metric of each stage.

C. Stage 3: Coarse-to-Fine Family Hierarchy

Because many confusions are between cheeses of the same kind, we group types into families. Crucially, family labels are *not* derived from the model's own embedding, since doing so is circular and inflates apparent accuracy. Instead, each cheese is matched by name to the cheese.com catalog and assigned a family from its curated `type`/`rind` field, mapped to eight families (*fresh, soft, bloomy-rind, washed-rind, semi-soft, semi-hard, hard, blue*). Name matching proceeds exact $\rightarrow$ substring/variant $\rightarrow$ fuzzy, resolving 769 of 1,238 types directly. The remainder falls back to a CLIP text-prototype assignment. We also define a coarser 5-family grouping through a map $\rho : \mathcal{F} \rightarrow \mathcal{F}'$ ($|\mathcal{F}'| = 5$) that merges the three soft families (*soft, bloomy-rind, washed-rind*) into one and the two semi families (*semi-soft, semi-hard*) into one, leaving *fresh*, *hard*, and *blue* unchanged. Family accuracy is obtained by *rolling up* the type model's predictions. Formally, let $\phi : \mathcal{C} \rightarrow \mathcal{F}$ map each of the $|\mathcal{C}| = 1{,}238$ types to a family. For a test image i with true type y_i and predicted type $\hat{y}_i$, family accuracy is

$$\mathrm{Acc}_{\mathcal{F}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\phi(\hat{y}_i) = \phi(y_i)], \tag{1}$$

i.e. a prediction counts as correct whenever it shares the true cheese's family, even if the exact type is wrong. The same form as the composed coarse map $\psi = \rho \circ \phi : \mathcal{C} \rightarrow \mathcal{F}'$ gives coarse-family accuracy.

Importantly, the external taxonomy is fixed before model evaluation and is not optimized using validation or test performance. The CLIP text-prototype fallback is used only when name alignment fails. It is fixed before training and does not use any test-set statistics or the recognition model's predictions on the evaluation data, so Stage 3 evaluation is not informed by the model under test.

D. Stage 4: Near-Duplicate Merge

We canonicalize each cheese to its base type using cheese.com names as anchors. The shortest catalog name that is a suffix of the folder name becomes the canonical label (so *"orkney extra mature cheddar"* $\rightarrow$ *cheddar*), with a stop-list so descriptors such as *gold* or *reserve* never become a base. Concretely, for a folder whose normalized, suffix-stripped token sequence is $t = (t_1, \ldots, t_n)$, with anchor set $\mathcal{A}$ (catalog names) and trailing-token frequencies $f(\cdot)$ over all folders, the canonical label is

$$c(t) = \begin{cases} s^{\star} & \text{if } \exists\, s \in \mathcal{A} \text{ a suffix of } t, \\ t_n & \text{else if } t_n \notin \mathcal{M},\ f(t_n) \geq \tau, \\ t & \text{otherwise,} \end{cases} \tag{2}$$

where $s^{\star}$ is the shortest such suffix, $\mathcal{M}$ is the modifier stop-list, and $\tau = 3$. This reduces 1,238 types to 1,033 (205 merged into 52 groups. The largest are *blue* (53 folders), *cheddar* (39), *brie* (14), *goat* (11), and *gouda* (10)). We then both (a) score the existing model under the merged labels and (b) retrain directly on the 1,033 classes, keeping the same image split.

The merging procedure is defined independently of model predictions and is based solely on string normalization and external catalog matching. No visual features or test set information are used in constructing merged labels.

V. Experiments and Results

All experiments use the fixed split (39,299 / 8,388 / 8,388 images) on a single NVIDIA RTX 4060 (8 GB). Stage-1 numbers are over the full archive. Stage-2–4 recognition numbers are held out for the test accuracy. We report cross-validation for the CLIP linear probe owing to its limited hyperparameter sensitivity. All other models are evaluated on the same fixed held-out test split.

Implementation details. Table II lists the fine-tuning configuration, identical across the backbone, merged, and class-balanced runs, so that only the studied variable changes.

TABLE II
FINE-TUNING CONFIGURATION (SHARED ACROSS RUNS).

Setting	Value
Backbone	ConvNeXt-Tiny / -Small (ImageNet-1k)
Input resolution	224×224
Batch size	48 (Small) / 64 (Tiny)
Optimizer	AdamW, lr 10^{-4}, weight decay 0.05
Schedule	cosine, up to 25 epochs, patience 4
Regularization	label smoothing 0.1, RandAugment$(2, 7)$
Precision	mixed (AMP)
Split (per cheese)	70 / 15 / 15 train/val/test
Hardware	1× NVIDIA RTX 4060 (8 GB)

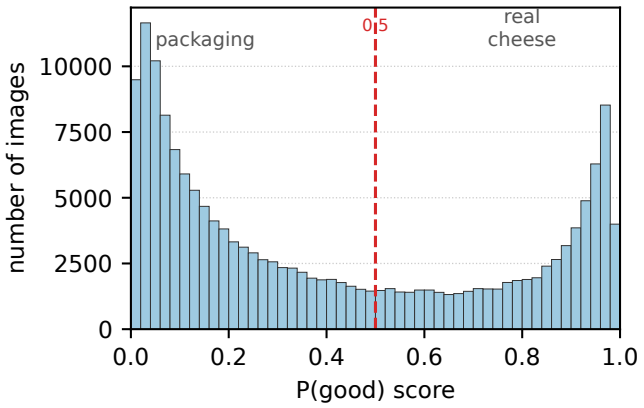

Fig. 3. Distribution of the packaging-filter score $P(\text{good})$ over all 165,769 images. The strongly bimodal shape (a packaging mode near 0 and a real-cheese mode near 1) shows why a 208-label linear probe separates the two classes. A few images sit near the 0.5 threshold.

TABLE III
STAGE 2: CHEESE-TYPE RECOGNITION (1,238 CLASSES).

Method	top-1	top-5
Uniform random baseline	0.08%	n/a
Zero-shot CLIP (text-name match)	4.4%	9.5%
CLIP linear probe (frozen, 3-fold CV)	36.7%	54.0%
ConvNeXt-Tiny fine-tuned (25 ep)	43.6%	66.9%
ConvNeXt-Small fine-tuned (25 ep)	**43.8%**	**68.0%**

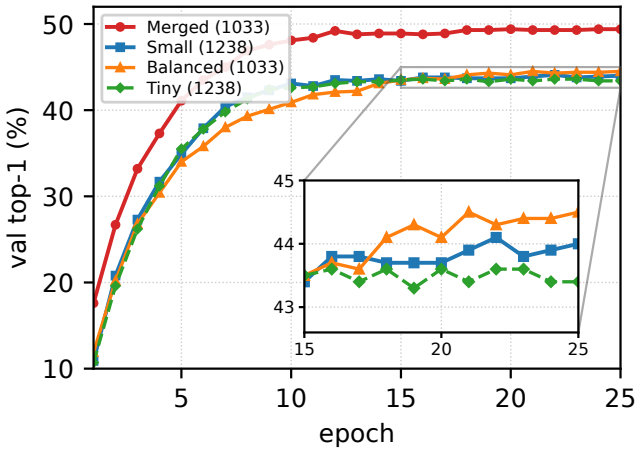

Fig. 4. Validation top-1 during fine-tuning. The larger backbone (Small) only slightly exceeds Tiny. Merging duplicate labels (Merged) raises the curve, while class-balanced training (Balanced) optimizes a different objective and trails in overall accuracy. The inset enlarges epochs 15–25, where Small, Balanced, and Tiny otherwise overlap, showing Balanced > Small > Tiny on the overall metric.

Mixed precision keeps peak memory near 5 GB at batch 48 for ConvNeXt-Small, well within the 8 GB budget. An epoch takes roughly three minutes, and runs early-stop on validation top-1 with patience four. CLIP features for Stages 1–3 are extracted once and cached.

A. Stage 1: Packaging Filter

Run over the full archive, the probe labels 62,174 images (37.5%) as real cheese and 103,595 as packaging or otherwise unusable. Because the classifier emits a calibrated score (the probability of the positive class), the operating point is adjustable. Raising the threshold yields a smaller but cleaner training pool, while the 0.4–0.6 band concentrates ambiguous cases for review. The score distribution is strongly bimodal (Fig. 3). Filtering is a prerequisite for the recognition stage, since training a type recognizer on the unfiltered folders would expose the model primarily to packaging rather than food.

B. Stage 2: Recognition

Table III compares methods. Zero-shot CLIP performs only slightly above a uniform random baseline (1,238-way, 0.08%), reflecting limited coverage of rare category names in the text encoder. A linear probe on the *same* frozen features reaches 36.7%, indicating that discriminative structure is present in the frozen image representations, and end-to-end fine-tuning adds a further ∼7 points of top-1 and ∼13 of top-5. Figure 4 shows validation top-1 over training.

C. Stage 3: Family Hierarchy

Accuracy rises sharply as labels coarsen. Against the exact-type baseline of 43.8% (1,238 classes), rolling the type model up to the authoritative 8-family taxonomy yields 64.7%, and the 5-family grouping yields 71.7%, so the model identifies the *kind* of cheese far more reliably than the exact type. A direct 8-family linear probe reaches only 46.2% (CV), lower than the roll-up because it must train on the noisier fallback labels. An earlier family map derived from CLIP prototypes scored higher (∼ 74%), but defining families from the model's own embedding is circular. The external taxonomy avoids this dependence. Figure 5 shows the family-level confusion of the type model on the test split (row-normalized). The diagonal dominates, per-family recall ranges from 57% to 76%, and off-diagonal mass concentrates between adjacent textures (soft / semi-soft / semi-hard / hard), exactly the boundaries that are genuinely ambiguous.

D. Stage 4: Merge, and Capacity vs. Data

Table IV reports the main comparison. Merging near-duplicate categories raises top-1 accuracy from 43.8% to

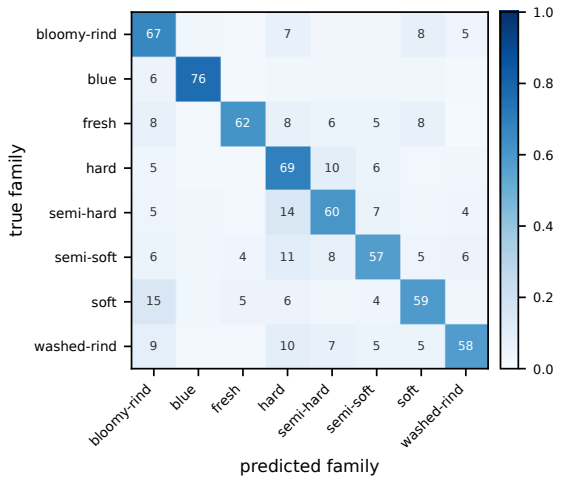

Fig. 5. Row-normalized family confusion matrix of the fine-tuned type model on the held-out test split (cell values are percentages of each true family). Errors are concentrated among neighboring textures rather than spread randomly.

TABLE IV
STAGE 4 AND LEVER ABLATION (IDENTICAL TEST SPLIT).

Setting	top-1	top-5
Baseline (Small, 1,238 raw labels)	43.8%	68.0%
+ bigger backbone (vs. Tiny, both 25 ep)	+0.2	+1.1
Scored by canonical label (relabel only)	48.6%	70.6%
Retrained on 1,033 merged classes	**49.0%**	**71.2%**
Merged + class-balanced loss	43.8%	68.6%

49.0% when the model is retrained on the merged label space, slightly above the 48.6% obtained by scoring the original predictions after mapping both the predicted argmax and the ground-truth labels to canonical classes, without retraining. The close agreement indicates that most of the gain comes from removing redundant labels rather than from additional training. By comparison, increasing capacity from ConvNeXt-Tiny to ConvNeXt-Small, both trained for 25 epochs, yields +0.2 points. Class-balanced loss reduces overall top-1 accuracy to 43.8%, since it down-weights frequent categories that also dominate the test distribution, trading overall accuracy for minority-class performance. These results indicate that, in this setting, accuracy is limited primarily by label-space structure rather than by model capacity. Figure 6 visualizes this progression.

E. Qualitative Results

Table V shows predictions from the fine-tuned model with the family rolled up by summing type probabilities within each family. The pattern matches the aggregate numbers. Visually distinct cheeses (mozzarella, cheddar) are identified exactly with high confidence, whereas a brie is taken for a camembert (a different named class but the same bloomy-rind product), and a Roquefort is ranked just below other blue cheeses. In every case, the *family* is correct, illustrating why the roll-up of Eq. (1) recovers so much accuracy. The model's mistakes are overwhelmingly within-family confusions among look-alikes.

TABLE V
QUALITATIVE PREDICTIONS (TOP-1 TYPE, FAMILY BY PROBABILITY SUM).

Query	Top-1 type (conf.)	Family (conf.)	Outcome
brie	mouco camembert (51%)	soft (60%)	near-twin
cheddar	cheddar (36%)	hard (85%)	exact
roquefort	blue vein (33%)	blue (63%)	family
mozzarella	mozzarella (95%)	semi-soft (96%)	exact

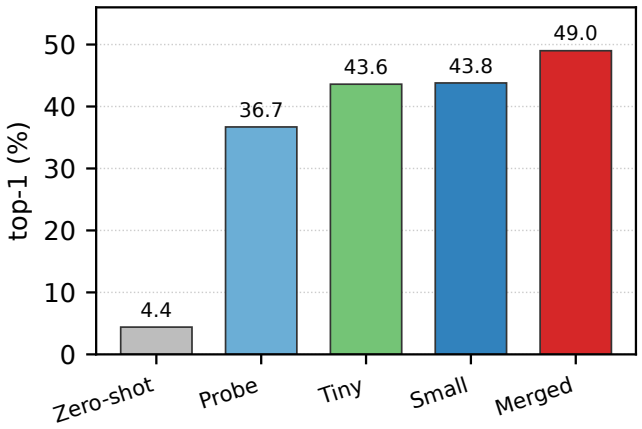

Fig. 6. Top-1 accuracy across the recognition progression. The jump from Small to Merged (label-space curation) exceeds the jump from Tiny to Small (model scaling), despite the latter doubling parameters and training time.

VI. DISCUSSION

Where the errors are. Manual inspection is consistent with the quantitative results. Errors are concentrated among visually similar categories rather than across distinct cheese types, and the model almost always returns the correct *family*. It cannot resolve genuine near-duplicates such as plain cheddar versus a named mature cheddar, which are not separable from the image content. The 43.8%→49.0% merge gain, therefore, corresponds largely to cases where the model identifies the correct cheese but is penalized for predicting a differently-named duplicate. This indicates that the primary limitation in this setting is ambiguity in label definitions rather than feature representation.

Capacity versus data. Once a competent backbone is fine-tuned, additional capacity and training time yield limited improvement, whereas curating the label space yields several points. For practitioners assembling web-scraped fine-grained corpora, this suggests prioritizing label hygiene (de-duplication, canonicalization, and a consistent taxonomy) before increasing model size. As summarized in Table IV, scaling the backbone is the most expensive intervention examined yet among the least effective (+0.2), de-duplication is inexpensive at inference time yet more effective (+5.2), and class balancing reduces accuracy on the overall metric. This does not imply that larger models are never beneficial. Rather,

under the evaluated model scales and dataset conditions, improvements from the label-space canonicalization exceed those obtained from moderate increases in backbone capacity.

Language grounding versus visual encoding. The gap between zero-shot CLIP (4.4%) and a linear probe on the *same* features (36.7%) indicates that the image encoder captures discriminative structure that a trained head can exploit, whereas the text encoder does not reliably ground rare proper nouns such as *"Abbaye de Belloc"* to appearance. The limitation lies primarily in language grounding rather than in visual encoding, which is why a small amount of supervision (a linear head, or the labels used in Stage 1) closes much of the gap.

Deployment. Because top-5 (71.2%) and coarse-family (71.7%) accuracy substantially exceed top-1, a practical system may present a short ranked list together with a family label rather than a single prediction.

Generality and future work. The approach is not specific to cheese. Corpora built by per-class web search tend to inherit the same three issues (off-target images, a large, noisy label vocabulary, and near-duplicate categories), so the four stages apply to web-scraped product, plant, or dish recognition wherever a domain catalog is available for the family and canonicalization maps. Possible extensions include using OCR of brand text on the discarded packaging images as a complementary signal, replacing the heuristic maps with verified references, and targeted collection of clean images for the sparsest canonical categories.

Limitations. The family and merge maps are partly heuristic and inherit gaps from the external catalog. The per-category image cap bounds tail coverage. And all results are obtained on a single archive and GPU, so the precise trade-off points may differ on other corpora, even if the qualitative conclusion holds.

VII. Conclusion

This study isolates the effect of label-space structure in a web-scraped fine-grained recognition setting. Across the evaluated configurations, label-space canonicalization yields larger performance gains than moderate increases in model capacity. Merging near-duplicate categories improves top-1 accuracy by 5.2 points, compared with 0.2 points from scaling ConvNeXt-Tiny to ConvNeXt-Small under an identical training budget.

The results suggest that in noisy, long-tailed datasets with redundant categories, improvements in data organization can yield larger gains than backbone scaling across the evaluated model scales. This conclusion is empirical, conditional on the dataset's structure and the range of model scales evaluated, and may not generalize to cleaner or substantially larger datasets.

References

[1] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proc. 38th Int. Conf. Machine Learning (ICML)*, Virtual Event, Jul. 2021, pp. 8748–8763.

[2] A. Dosovitskiy *et al.*, "An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale," in *Proc. Int. Conf. Learning Representations (ICLR)*, Virtual Event, 2021.

[3] G. Ilharco *et al.*, "OpenCLIP," Zenodo, Jul. 2021. doi: 10.5281/zenodo.5143773.

[4] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," California Institute of Technology, Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.

[5] G. Van Horn *et al.*, "The iNaturalist Species Classification and Detection Dataset," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 8769–8778.

[6] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101: Mining Discriminative Components With Random Forests," in *Proc. Eur. Conf. Computer Vision (ECCV)*, Zurich, Switzerland, Sep. 2014, pp. 446–461.

[7] Z. Liu *et al.*, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 11966–11976.

[8] W. Li *et al.*, "WebVision Database: Visual Learning and Understanding From Web Data," arXiv preprint arXiv:1708.02862, Aug. 2017.

[9] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-Balanced Loss Based on Effective Number of Samples," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 9268–9277.

[10] M. Ache, "Cheese Pics," Kaggle Dataset, 2021. [Online]. Available: https://www.kaggle.com/datasets/mathurinache/cheese-pics

[11] TidyTuesday, "Cheeses Dataset," Jun. 4, 2024. [Online]. Available: https://github.com/rfordatascience/tidytuesday. [Accessed: Jun. 2, 2026].

[12] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *Proc. Int. Conf. Learning Representations (ICLR)*, New Orleans, LA, USA, 2019.

[13] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "RandAugment: Practical Automated Data Augmentation With a Reduced Search Space," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA, Jun. 2020, pp. 3008–3017.