MeowWoofNet: A Hybrid Semi-Supervised and Masked Fine-Tuning Framework for Pet Recognition

Ville Sebastian Olsson seo@kth.se **Silpa Soni Nallacheruvu** ssnal@kth.se **Raahitya Botta** raahitya@kth.se **Joakim Axnér** jaxner@kth.se

Abstract

Deep learning for fine-grained image classification typically demands both extensive labeled data and expensive full-model fine-tuning. In this work, we present MeowWoofNet, a hybrid framework that combines FixMatch semi-supervised learning (SSL) with Gradient-based Parameter Selection (GPS) to address these bottlenecks in the Oxford-IIIT Pet Dataset. The FixMatch approach is employed to leverage unlabeled data in scenarios with limited labeled supervision, demonstrating that comparable 91.74% accuracy to fully supervised models can be achieved with as little as 50% labeled data. A systematic analysis of pseudo-label confidence thresholds reveals important trade-offs between label quality and supervisory signal quantity. Furthermore, GPS-based masked fine-tuning is introduced to significantly reduce the number of trainable parameters, yielding competitive performance while improving computational efficiency. Finally, integrating GPS into the FixMatch loop yields 92.19 ± 0.11 % accuracy—only 0.73 percentage points below GPS alone—validating compatibility. Our results indicate that MeowWoofNet attains near–state-of-the-art performance with minimal annotation and compute, offering a practical path for resource- efficient transfer learning.

1 Introduction

Deep learning has driven state-of-the-art advances in computer vision, but its success often hinges on two costly resources: large labeled datasets and expensive full-model fine-tuning. Semi-supervised learning (SSL) mitigates the data bottleneck by leveraging abundant unlabeled examples alongside scarce labels. Among SSL methods, we chose FixMatch over MixMatch or Mean Teacher due to its conceptual simplicity, strong empirical results, and intuitive use of weak and strong augmentations: it generates pseudo-labels from weakly augmented images when the model's confidence exceeds a threshold, then enforces consistency under strong augmentations, which suited the high intra-class variability of pet images. Meanwhile, parameter-efficient transfer learning (PEFT) techniques seek to reduce the cost of adapting large pre-trained models to new tasks by updating only a small subset of parameters. Gradient-based Parameter Selection (GPS) exemplifies a non-destructive, model-agnostic PEFT approach: it ranks parameters by gradient magnitude on a downstream task, masks out the lowest-impact weights, and fine-tunes only a top fraction of parameters.

Our research question explores whether the integration of semi-supervised learning and masked fine-tuning can substantially reduce labeling effort and computational overhead, while preserving model accuracy in the context of pet recognition. First, we implement FixMatch to study how varying the pseudo-label confidence threshold affects performance under limited labels. Second, we reproduce GPS-style masked fine-tuning to assess whether sparse parameter updates can reliably outperform full fine-tuning, examining robustness to random seeds and hyperparameter settings within a reasonable training time. By combining these strategies, we have built MeowWoofNet, a hybrid semi-supervised and masked fine-tuning framework that achieves high accuracy for pet recognition with minimal annotation and compute overhead.

2 Related work

Semi-supervised learning has emerged as a powerful approach for enhancing model performance in scenarios with limited labeled data. Within this domain, the FixMatch algorithm *Sohn et al., 2020* [1] stands out as an elegant yet remarkably effective solution that seamlessly integrates pseudo-labeling with consistency regularization techniques. The core mechanism of FixMatch involves generating pseudo-labels from weakly augmented unlabeled images where the model produces high-confidence predictions, then enforcing consistent predictions when those same images undergo strong augmentation. This straightforward approach effectively leverages unlabeled data without necessitating complex architectural modifications or elaborate processing pipelines, achieving state-of-the-art performance across a variety of standard semi-supervised learning benchmarks, including 94.93% accuracy on CIFAR-10 with 250 labels.

In our project, we implemented the FixMatch framework to address the challenge of limited labeled data when working with the Oxford-IIIT Pet Dataset. Building upon the foundational work of *Sohn et al.* and drawing inspiration from *Zhang et al. (2023)* [2], we conducted a systematic investigation into the importance of dynamic thresholds for selecting trustworthy pseudo-labels for the unlabeled data. Specifically, we analyzed the critical trade-off between pseudo-label quality and the amount of supervisory signal obtained from unlabeled data by varying the pseudo-label confidence threshold globally across all classes. This contrasts with the per-class pseudo-label screening ratio approach used in the DTA algorithm presented by *Zhang et al. (2023)* [2], allowing us to explore a simpler yet effective alternative for pseudo-label filtering.

Recent efforts in parameter-efficient transfer learning have explored *masked fine-tuning*, in which only a subset of model parameters is updated. While SSL tackles the data bottleneck by leveraging unlabeled images, masked fine-tuning addresses the complementary challenge of reducing adaptation cost when transferring large pre-trained networks to new tasks. Gradient-based Parameter Selection (GPS) by Zhang *et al.* (2024) first introduced this approach, identifying important parameters via gradient magnitude and fine-tuning only those weights [3]. On image classification benchmarks (FGVC, VTAB), GPS updated merely $\sim 0.36\%$ of a pre-trained model's parameters—yet matched or exceeded full fine-tuning accuracy. On the Oxford IIIT Pet dataset, GPS achieved a 91.7 % test accuracy—a substantial improvement compared to full model fine-tuning (86.9 %). Unlike adapter- or LoRA-based methods, GPS adds no extra modules: it simply ”masks” (freezes) most weights and optimizes a small, gradient-selected subset, demonstrating that extreme sparsity can preserve—or even improve—task performance at a fraction of the computational cost.

3 Data

Our project is based on the Oxford-IIIT Pet Dataset [4], a widely used benchmark for fine-grained image classification. The dataset contains 7,349 images of cats and dogs spanning 37 different breeds, with roughly 200 images per class. Each image is annotated with both a category label (breed) and a binary label (cat vs. dog).

Following standard practice, we use the `trainval` and `test` splits defined by the dataset authors, with further internal division of `trainval` into training, validation, and unlabeled subsets for different experiments. We apply basic pre-processing, such as resizing, normalization, and optional horizontal flipping, during training. For our semi-supervised experiments, we implement the FixMatch framework, which applies weak augmentation involves random cropping and horizontal flipping, and strong augmentation includes color jitter, grayscale conversion, and random rotations to unlabeled data.

The Oxford-IIIT Pet dataset is considered challenging [4] due to its small size, fine-grained labels, and large intra-class variance. On the full 37-class task, pretrained ResNet-50 models typically achieve 88–91% top-1 accuracy with standard fine-tuning [5]. With sharpness-aware minimization (SAM) and heavy data augmentation, top ResNet-50-based methods reach up to ∼95% [6], though such results often rely on advanced tuning or ensembling. Larger architectures like Vision Transformers (ViTs) achieve higher performance: DINOv2 (ViT-B/14, frozen) reaches 96.7% [7], and OmniVec2, a large multimodal transformer, reports state-of-the-art performance at 99.6% accuracy [8]. Since our experiments focus on fine-tuning ResNet backbones without massive pretraining or scale, we consider 90–94% test accuracy a strong result in our setting.

4 Methods

The Top-1 accuracy metric was used throughout this paper to assess model performance. All metrics were measured on a single NVIDIA L4 GPU. ResNet-50 was chosen over smaller variants (e.g., ResNet-18) due to its higher performance with less tuning effort.

4.1 Baseline Fine-Tuning

To test our training pipeline, we began with a binary classification task (Dog vs. Cat) using a pretrained ResNet-50. We replaced the final classification layer and fine-tuned it on the Oxford-IIIT Pet Dataset. We used the Adam optimizer with a learning rate of 10^{-3} and no weight decay, as only the final layer was trained. Epoch count was selected based on validation performance.

Extending to multi-class classification (37 breeds), we compared two fine-tuning strategies: (1) simultaneous unfreezing of the last l layers from the start of training, and (2) gradual unfreezing, where earlier layers are unfrozen at predefined epochs. We applied standard augmentations (random flips and rotations), and used a ReduceLROnPlateau scheduler. To find the best hyperparameters for both strategies, we used Optuna (a hyperparameter optimization framework) to perform an efficient search over learning rate, weight decay, momentum, dropout rate, and unfreezing epochs. The optimal parameters were then reused across experiments to ensure a fair comparison.

The effects of class imbalance on test accuracy was investigated by downsampling all cat breeds to 20% of their original size while keeping dog breeds intact. ResNet-50 was then fine-tuned under four imbalance handling configurations: (i) unmodified cross-entropy (CE), (ii) weighted CE (cat samples weighted by 5), (iii) oversampling (higher sampling probability for cats), and (iv) both weighting and oversampling. All settings used the binary classification hyperparameters.

4.2 Semi-supervised Learning

This section describes the FixMatch algorithm, a semi-supervised learning strategy inspired by *Sohn et al.* [1]. The model's predictions on weakly augmented images are treated as pseudo-labels, and consistency is enforced by training the model to predict the same labels on strongly augmented versions, enhancing robustness and generalization.

To maintain the quality of these pseudo-labels, a confidence threshold τ is optionally enforced—only high-confidence predictions are used for supervision. The unsupervised loss from these pseudo-labeled samples is scaled by a weighting factor λ, balancing their influence with the labeled data. The FixMatch loss function used in our method is defined as $\mathcal{L} = \mathcal{L}_s + \lambda \times \mathcal{L}_u$, where:

$$\mathcal{L}_s = \frac{1}{B}\sum_{b=1}^{B} \text{CE}\left(y_b, p_\theta\left(y \mid \mathcal{A}_{\text{weak}}(\mathbf{x}_b)\right)\right), \mathcal{L}_u = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\left[\max(\hat{y}_b) \geq \tau\right] \cdot \text{CE}\left(\hat{y}_b, p_\theta\left(y \mid \mathcal{A}_{\text{strong}}(\mathbf{u}_b)\right)\right)$$

Here, $\hat{y}_b$ denotes the pseudo-label assigned to an unlabeled sample $\mathbf{u}_b$, while the partitioning into supervised and unsupervised datasets is governed by the labeled data fraction l. The confidence threshold τ was set to the default value of 0.95 throughout our experiments.

4.3 Masked fine-tuning

Following the Gradient-based Parameter Selection (GPS) paradigm [3], we implement masked fine-tuning as a two-stage, parameter-efficient transfer-learning method. In the first stage, we loop over the entire labeled training set once, accumulating the absolute gradient of every backbone weight. For each neuron, we sort its incoming weights by descending gradient magnitude and retain only the top-K weights, yielding a sparse binary mask $M \in \{0, 1\}^{\#\text{weights}}$. By varying K, we control the fraction of tunable parameters.

In the second stage, we freeze all weights where $M = 0$ and only allow updates to the selected subnetwork. We combine this with gradual layer unfreezing. During each optimization step, gradients on the unselected weights are zeroed out before applying NAG updates using our standard learning-rate, momentum, and weight-decay hyperparameters. This ensures all unselected weights remain at their pre-trained values, incurring no extra inference costs while reducing the number of tunable parameters by orders of magnitude.

5 Experiments

5.1 Baseline Fine-Tuning

For the binary Dog vs. Cat classification task, we used a ResNet-50 with frozen convolutional layers and Adam optimizer (10^{-3} learning rate, no weight decay). We stopped trained after 3 epochs. The model achieved a mean test accuracy of 99.33% across five runs, with a standard error of 0.02 percentage points.

For multi-class classification (37 breeds), we compared two fine-tuning strategies on ResNet-50: (1) simultaneous unfreezing of the last l layers (Strategy 1), and (2) gradual unfreezing of earlier layers during training (Strategy 2). Hyperparameters were optimized using Optuna, and we used NAG with a ReduceLROnPlateau scheduler. Strategy 1 consistently outperformed Strategy 2, as shown in Table 1. The best result was 93.47% test accuracy when unfreezing layers 1–4, while including `conv1` slightly decreased performance for both strategies, suggesting diminishing returns from early-layer tuning.

Fine-tuned layers	**Strategy 1**	**Strategy 2**
Layer 4 & 3	93.39 ± 0.01	92.89 ± 0.08
Layer 4, 3 & 2	93.36 ± 0.10	92.98 ± 0.04
Layer 4, 3, 2 & 1	93.47 ± 0.01	93.04 ± 0.04
Layer 4, 3, 2, 1 & `conv1`	93.27 ± 0.03	92.91 ± 0.11

Table 1: Comparison of mean test accuracy (%) and standard errors between Strategy 1 and 2.

To investigate the effects of class imbalance, we ran each configuration three times (seeds 42, 43, 44) and computed the mean test accuracy and standard error. We observed that both oversampling and loss reweighting improved test accuracy under class imbalance, with the combination of both yielding the best performance. Oversampling alone increased accuracy by 3.4 percentage points, and combining it with weighted loss produced the best result at 87.57%. See Table 6.

5.2 Semi-supervised learning

To determine the optimal unsupervised loss weight λ in our FixMatch implementation, we varied λ from 0.1 to 1.0 in increments of 0.1. While prior work such as [1] used fixed numbers of unlabeled samples and larger λ values (e.g., $\lambda = 10$) to emphasize their influence, our setup splits the dataset into labeled and unlabeled fractions ranging from 0.1 to 1.0. Thus, we explored a moderate λ range to reflect the proportional increase in unlabeled data. Experiments were conducted with 0.5 labeled fraction over 5 epochs, using fixed hyperparameters from our tuning procedure (see subsection 8.2). The results are summarized in Table 2.

λ	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
Test Accuracy (%)	89.94	89.97	89.94	89.94	90.00	90.00	89.94	89.94	89.97	89.86
Final Loss	0.4033	0.4181	0.4329	0.4476	0.4624	0.4771	0.4919	0.5067	0.5217	0.5366

Table 2: Comparison of final test accuracy (%) and loss for different λ values with seed 42.

As shown in Table 2, since the test accuracy remains stable across all values of λ, with the highest accuracy achieved at $\lambda = 0.5$ and 0.6 at 5 epochs already, we select $\lambda = 0.5$ as the optimal balance between accuracy and loss for the following runs. To evaluate FixMatch under varying supervision levels, we ran experiments with labeled data fractions of 0.1 to 1.0, using the rest as unlabeled. Test accuracy was measured after 10, 20, and 30 epochs and averaged over seeds 42, 43, and 44 for robustness, with results summarized in Table 3.

Epochs	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**	**1.0**
10	84.90	90.79	91.74	92.02	92.47	92.84
20	84.54	89.96	91.14	91.82	92.07	93.01
30	83.12	89.61	90.30	91.58	92.23	92.84

Table 3: Comparison of mean test accuracy (%) for different labelled data fractions and epochs.

As reflected in Table 3, semi-supervised learning exhibits its most significant advantage at lower labeled fractions, where the FixMatch approach attains a remarkable 84.90% accuracy using merely 10% of the labeled data. Model performance generally peaks around 10 epochs, with a decline observed at 30 epochs, suggesting the onset of overfitting in the absence of sufficient regularization. Notably, the marginal gains diminish as more labeled data is introduced: a substantial improvement is seen when increasing the labeled fraction from 0.1 to 0.3, whereas the jump from 0.7 to 0.9 offers only minimal benefit. This pattern remains consistent across all training durations, underscoring the point of diminishing returns. Remarkably, the model trained on just 50% of the labeled data performs on par with the fully supervised counterpart, affirming the practical efficacy of semi-supervised learning in resource-constrained scenarios.

Inspired by *Zhang et al. (2023)*, we investigated how different pseudo-labeling thresholds affect model performance. To prevent discarding correct pseudo-labels due to a fixed high threshold in FixMatch [2], we experimented with thresholds of 0.5, 0.7 and disabled. To test whether the impact of pseudo-labeling is typically more pronounced when labeled data is scarce, we evaluated the mean test accuracy across various labeled data fractions summarized below in Table 4.

Threshold	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
None (Disabled)	85.04	90.73	91.65	91.99	92.02
0.5	83.80	89.77	91.51	91.88	92.25
0.7	83.74	89.78	91.13	91.68	91.79

Table 4: Comparison of mean test accuracy (%) for different labeled data fractions and τ.

Disabling the confidence threshold proves beneficial in low-resource settings, where accepting a larger pool of pseudo-labels—even if noisy—helps compensate for limited supervision. Conversely, when labeled data is more abundant, maintaining the default threshold yields better results by preserving label quality. This trade-off between label quantity and reliability underscores the importance of adapting pseudo-labeling strategies based on data availability. To better capture these dynamics beyond early training behavior, we report performance at 20 epochs rather than stopping at 10. Across the experiments in Table 3 and Table 4, the standard error in test accuracy remained within 0.03–0.5 percentage points, supporting the statistical reliability of our results.

5.3 Masked fine-tuning

GPS has been found to reduce per-mini-batch training time from ~ 140 ms to ~ 110 ms compared to full fine-tuning [3]. To reproduce this result, the network was trained with all layers unfrozen, with gradual unfreezing, and with gradual unfreezing combined with masked fine tuning using sparsity levels $K \in \{10, 100, 1000, 4607\}$. Here, $K = 4607$ is a boundary value in the sense that it is the maximum value which retains at least one zero entry in the mask. The elapsed time of training for 10 epochs was recorded and each run was repeated $N = 3$ times with different seeds.

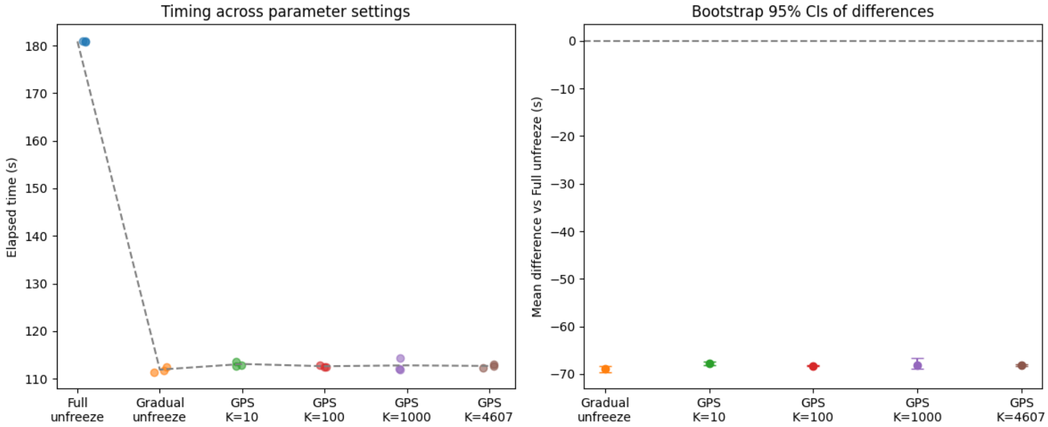

Figure 1: **Left:** Jitter plot of the total training time for each freezing strategy. **Right:** Mean difference in seconds relative to the full-unfreeze baseline along with confidence intervals.

Figure 1 shows that both gradual unfreezing and gradual-unfreeze + GPS finish in ~ 110 s, shaving ~ 70 s off full unfreeze. This is consistent with the earlier finding, but also suggests that the speedup in this case stems entirely from layer freezing and not from masking overhead.

In the second experiment, we investigate the finding that GPS may improve accuracy. Using gradual unfreezing as baseline, we do this by checking how the test accuracy differs when applying masked fine-tuning for different values of K. Finally, we investigate whether we can combine masked fine-tuning with SSL and still retain high accuracy. We do this by repeating the experiment under FixMatch with 50 % labeled data, $\lambda = 0.5$, and a pseudo-threshold of 95 %. The results are summarized in Table 5.

Model	**No GPS**	K=10 (0.19%)	K=100 (0.21%)	K=1000 (62.56%)	K=4607 (99.98%)
No SSL	92.66 ± 0.09	53.79 ± 0.42	83.61 ± 0.07	92.92 ± 0.10	92.66 ± 0.05
SSL	92.24 ± 0.12	49.52 ± 0.28	77.60 ± 0.15	92.19 ± 0.11	92.28 ± 0.12

Table 5: Mean test accuracy (%) and standard deviation for different GPS parameter settings with and without SSL. The percentages for K indicate the fraction of parameters that are masked.

At very low K (10 or 100), the model severely underfits, dropping to 54 % or 84 % accuracy—well below both full and gradual unfreeze. Once you hit $K = 1000$, we cross a "sparsity threshold" where the selected subnetwork is large enough to capture the key downstream features. Here we not only recover the 92.66% baseline but actually see a slight uptick to 92.92%—suggesting that pruning out the least-useful weights may even regularize the model a bit. At $K = 4607$, we exactly match the baseline again, confirming that there is diminishing return beyond a certain subnetwork size. Upon adding SSL for $K = 1000$, the accuracy dropped by only 0.73 percentage points.

6 Conclusion

This project set out to explore the effectiveness of semi-supervised learning in environments constrained by limited labeled data. Leveraging the FixMatch framework, we examined how a combination of pseudo-labeling and consistency regularization can be harnessed to train high-performing models with minimal annotation overhead. Our investigations confirmed that when guided by a well-balanced unsupervised loss weight, unlabeled data becomes a powerful supervisory signal—allowing the model to learn robust representations without being overwhelmed by noise from imperfect pseudo-labels.

We found that FixMatch is not only scalable and lightweight in implementation but also remarkably resilient across varying degrees of supervision. In particular, the model was able to achieve performance competitive with fully supervised baselines even when trained on a fraction of labeled data. Our experiments further revealed that the effectiveness of pseudo-labeling can be enhanced by adaptively relaxing the confidence threshold, especially in low-resource settings where strict filtering may prematurely discard useful supervisory signals.

Beyond semi-supervised learning, we explored the emerging paradigm of masked fine-tuning, where only a sparse subset of network parameters are updated based on gradient importance. Our findings support the conclusion that a carefully selected subnetwork can be sufficient to fine-tune the model effectively. When combined with FixMatch, this masked approach retained impressive accuracy while significantly reducing the computational footprint. Notably, even under the constraints of gradual unfreezing, the sparse adaptation strategy yielded performance nearly indistinguishable from full fine-tuning, confirming that the benefits of parameter efficiency and data efficiency are not mutually exclusive. As future work, it would be worth investigating whether mask sizes between 100 and 1000 could achieve similarly strong results or further optimize the trade-off between sparsity and accuracy.

Together, these results underscore a compelling vision for future learning systems: ones that are not only label-efficient but also compute-efficient, capable of scaling to real-world problems with limited supervision and constrained resources. Looking forward, these insights invite further exploration into adaptive thresholding strategies, task-aware masking techniques, and their deployment in domains such as personalized healthcare, autonomous systems, and low-shot recognition tasks—settings where annotation costs and model size are both critical factors.

7 References

[1] Kihyuk Sohn et al. "FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence". In: *arXiv preprint arXiv:2001.07685* (2020).

[2] Tianyou Zhang et al. "Semi-supervised learning for classification of radio galaxies using consistency regularization and pseudo-labeling". In: *Research in Astronomy and Astrophysics* 23.11 (2023), p. 125501.

[3] X. Zhang, Y. Liu, and et al. "Gradient-based Parameter Selection for Masked Fine-Tuning". In: *arXiv preprint arXiv:2312.10136* (2024).

[4] Omkar M Parkhi et al. "Cats and dogs". In: *CVPR* (2012).

[5] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. "When Vision Transformers Outperform ResNets without Pre-training or Strong Data Augmentations". In: *Proc. of ICLR*. 2022.

[6] Feng Wang et al. "Self-Supervised Learning by Estimating Twin Class Distributions". In: *NeurIPS Workshops (Datasets and Benchmarks Track)*. 2021.

[7] Maxime Oquab et al. "DINOv2: Learning Robust Visual Features without Supervision". In: *Trans. Mach. Learn. Res.* (2024).

[8] Siddharth Srivastava and Gaurav Sharma. "OmniVec2 – A Novel Transformer based Network for Large Scale Multimodal and Multitask Learning". In: *Proc. of CVPR*. 2024.

8 Appendix

8.1 Project code

The project code for MeowWoofNet can be found at:

https://github.com/Sebelino/DD2424-project

8.2 Hyperparameters for SSL

Experiments of semi-supervised learning used the following hyperparameters from the hyperparameter search: learning rate $\eta = 0.0047$, momentum $= 0.7222$, weight decay $= 4.2874 \times 10^{-6}$. We used the NAG optimizer for enhanced performance, along with our fine-tuning strategy that combined Strategy 1 (unfreezing the last two layers) and Strategy 2 (gradual unfreezing at epochs 2 and 4), as outlined in subsection 5.1.

8.3 Imbalanced Classes Results

To illustrate how class imbalance affects per-class accuracy, we include a confusion matrix in Figure 2. The matrix displays absolute counts rather than percentages. Correct predictions appear along the diagonal (top-left to bottom-right), while off-diagonal entries indicate misclassifications. We observe that cat classes are misclassified more frequently than dog classes, and that misclassified cats are more often confused with other cat breeds than with dogs.

Final test accuracies are reported in Table 6.

Imbalance Strategy	**Test Acc. (%)**
Standard CE	83.88 ± 0.22
Weighted CE	85.11 ± 0.25
Oversampling	87.31 ± 0.17
Weighted + Oversampling	87.57 ± 0.20

Table 6: Final test accuracy under different imbalance handling strategies (3 runs).

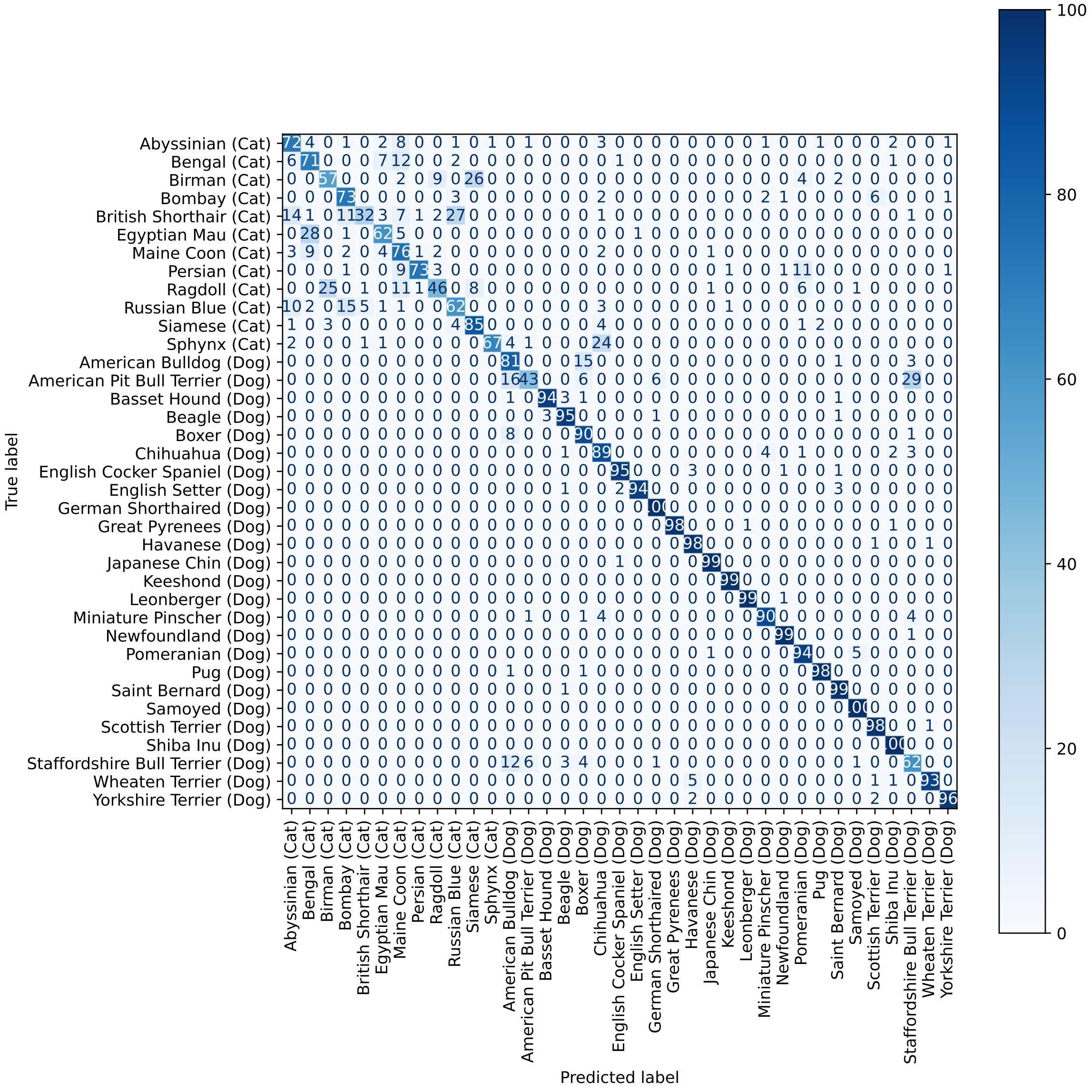

Figure 2: Confusion matrix for multi-class classification, ordered with cat breeds first.