COMPUTE-EFFICIENT FINE-TUNING OF RESNET-18 FOR FINE-GRAINED FOOD IMAGE CLASSIFICATION

Mario Aldahir Escalante-Contreras
Business Intelligence / Data Visualization (Data Science)
Montreal College of Information Technology
Montreal, Canada

ABSTRACT

We study fine-tuning strategies for food image classification using the Food-101 dataset. Our objective is to build a robust classifier under realistic compute constraints (Google Colab GPU sessions) while maintaining strong generalization. We fine-tuned a pretrained ResNet-18 model by freezing early layers and training higher-level residual blocks (layers 2–4) with a lightweight classification head. To improve generalization, we incorporate label smoothing, MixUp augmentation, cosine learning-rate scheduling, mixed-precision training (AMP), and gradient clipping. We further evaluate a two-phase resolution strategy: training in 192×192 and optional final fine-tuning of "polish" in 224×224. Our best model achieves a top-1 validation accuracy of 78.5% after the 224×224 phase, improving over the 192×192 model (77.1%). We present empirical evidence showing the impact of selective freezing strategies, and we analyze common failure cases (visually similar dishes, occlusion, and background bias). Our results demonstrate that careful fine-tuning and augmentation can yield strong performance with modest model capacity and limited compute.

1 INTRODUCTION

Food image recognition has applications in dietary logging, restaurant search, health monitoring, and assistive technologies. However, food classification remains challenging due to high intra-class variability (different presentations of the same dish) and inter-class similarity (visually similar dishes). Although large models can achieve strong performance, many practitioners operate under constrained compute environments such as Google Colab Google Research (2023), where session limits and I/O bottlenecks can dominate training cost.

In this work, our objective is to build an accurate Food-101 classifier Bossard et al. (2014) using transfer learning Pan & Yang (2010) and compute-efficient fine-tuning. We focus on practical training strategies, including selectively unfreezing higher-level residual blocks, applying strong regularization techniques (MixUp Zhang et al. (2018) and label smoothing Szegedy et al. (2016)), and performing a final high-resolution fine-tuning stage to recover fine-grained texture details. Our approach achieves a final top-1 validation accuracy of 78.5% while keeping training feasible under limited compute budgets.

2 PROBLEM STATEMENT

We consider the task of supervised food image classification using the Food-101 dataset Bossard et al. (2014). Given an RGB food image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$, the objective is to predict a categorical label

$$y \in \{1, 2, \ldots, 101\},$$

corresponding to one of the predefined food categories.

Let $f_\theta : \mathbb{R}^{H \times W \times 3} \to \mathbb{R}^{101}$ denote a deep neural network parameterized by θ, which outputs a vector of class logits. The predicted label is obtained via

$$\hat{y} = \arg\max_k \; f_\theta(\mathbf{x})_k.$$

The model is trained to minimize the empirical risk over a labeled training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ using a cross-entropy-based objective:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell\big(f_\theta(\mathbf{x}_i), y_i\big),$$

where $\ell(\cdot)$ denotes the categorical cross-entropy loss, optionally augmented with regularization strategies such as label smoothing.

We evaluate the learned classifier using top-1 classification accuracy on a held-out validation split. The primary goal of this work is to maximize validation accuracy while maintaining practical training time, memory efficiency, and stability within constrained cloud-based environments such as Google Colab. These constraints motivate the use of transfer learning, selective fine-tuning of network layers, and compute-efficient optimization strategies.

3 Background and Related Work

Convolutional neural networks (CNNs) remain strong baselines for image classification Krizhevsky et al. (2012), with residual networks enabling deeper architectures through the use of skip connections He et al. (2016). In particular, ResNet-18 is widely adopted due to its favorable trade-off between accuracy and computational efficiency He et al. (2016). Transfer learning further improves practicality by leveraging pretrained weights, commonly from ImageNet Deng et al. (2009), to reduce data requirements and accelerate convergence when adapting models to new domains Pan & Yang (2010). A standard fine-tuning strategy consists of freezing early layers that capture generic visual features while training higher-level layers to adapt semantic representations to the target dataset.

Data augmentation and regularization play a critical role in improving generalization. Label smoothing has been shown to reduce over-confidence and improve model calibration Szegedy et al. (2016), while MixUp constructs convex combinations of training samples and their corresponding labels, often yielding improved robustness Zhang et al. (2018). Learning-rate scheduling strategies such as cosine annealing are commonly employed to stabilize optimization during fine-tuning Loshchilov & Hutter (2017). Our approach integrates these established techniques to maximize performance under constrained training environments.

4 Method

4.1 Method Overview

We adopt a transfer learning approach based on a pretrained convolutional neural network, leveraging strong data augmentation and selective fine-tuning to achieve high classification accuracy under limited computational resources. Our pipeline is built around a ResNet-18 backbone He et al. (2016) pretrained on ImageNet Deng et al. (2009), followed by a lightweight classification head tailored to the Food-101 task.

Training proceeds in two stages. First, the model is trained using images resized to 192×192 pixels with aggressive data augmentation and partial unfreezing of higher-level backbone layers. This stage allows the model to adapt semantic features to the food domain while preserving general visual representations learned from ImageNet. Second, we optionally perform a short fine-tuning phase at a higher input resolution of 224×224 pixels to recover fine-grained texture and shape details that may be lost at lower resolution. Across both stages, we employ mixed-precision training, cosine learning-rate scheduling, label smoothing, and MixUp regularization to improve generalization and training stability.

4.2 Data Preprocessing

Input images are decoded as RGB tensors and normalized using the standard ImageNet mean and standard deviation:

$$\mu = (0.485, 0.456, 0.406), \quad \sigma = (0.229, 0.224, 0.225).$$

During training, we apply a sequence of stochastic augmentations designed to increase robustness to variations in scale, viewpoint, and illumination. These include *RandomResizedCrop* with scale and aspect ratio jitter, horizontal flipping, *ColorJitter* for brightness, contrast, and saturation changes, and *RandomErasing* Zhong et al. (2020) to simulate occlusions and reduce overfitting.

For validation, we use deterministic preprocessing consisting of resizing followed by center cropping to ensure consistent evaluation. No stochastic augmentations are applied at validation time.

4.3 Model Architecture

Our model uses ResNet-18 He et al. (2016) as the feature extractor. ResNet-18 is a lightweight residual network that balances representational capacity and computational efficiency, making it suitable for cloud-based training environments.

The original ImageNet classification layer is replaced with a custom head consisting of two fully connected layers:

$$512 \rightarrow 256 \rightarrow 101,$$

where 101 corresponds to the number of Food-101 classes. A ReLU activation follows the intermediate layer, and dropout is applied to mitigate overfitting. This design provides additional task-specific capacity while keeping the parameter count modest.

4.4 Fine-Tuning Strategy

To satisfy the instructional requirement of selective fine-tuning, we freeze the early layers of the network, including the initial convolutional layer and the first residual block (layer1). These layers primarily capture low-level visual features such as edges and textures, which are generally transferable across domains.

We fine-tune the higher-level residual blocks (layer2, layer3, and layer4) along with the classification head. This strategy allows the model to adapt mid-level and semantic features to the food classification task while maintaining training stability. Preliminary experiments training only the final block and classifier were conducted, but unfreezing layer2–layer4 consistently yielded superior validation accuracy and was therefore adopted in the final configuration.

4.5 Optimization Details

The model is trained using the categorical cross-entropy loss with label smoothing Szegedy et al. (2016), which mitigates overconfident predictions and improves generalization. MixUp Zhang et al. (2018) augmentation is applied during training with mixing coefficient $\alpha = 0.2$, while validation is performed without MixUp.

Optimization is performed using the AdamW optimizer Loshchilov & Hutter (2019) with layer-wise learning rates, assigning higher learning rates to the classifier head and progressively smaller rates to deeper backbone layers. A cosine annealing learning-rate schedule Loshchilov & Hutter (2017) is used to gradually reduce the learning rate over the course of training.

To improve numerical stability and efficiency, we employ automatic mixed-precision training Micikevicius et al. (2018), using bfloat16 or float16 depending on hardware support. Gradients are clipped to a maximum ℓ_2 norm of 1.0 to prevent instability during fine-tuning. Additionally, tensors are stored in `channels_last` memory format to improve convolution performance on modern GPUs.

Due to session time limits and possible interruptions in Google Colab, we implement robust checkpointing that saves both the latest model state and the best-performing model according to validation accuracy, enabling seamless training resumption.

5 Experiments and Methodology

This section describes the dataset, experimental setup, evaluation metrics, and baseline comparisons used to assess the proposed approach. All experiments are designed to be reproducible and to reflect practical training constraints in a cloud notebook environment.

5.1 DATASET

We evaluate our method on the **Food-101** dataset Bossard et al. (2014), a large-scale food image classification benchmark consisting of 101 food categories with 1,000 images per class, for a total of 101,000 RGB images. The dataset exhibits significant intra-class variability due to diverse food presentations, as well as inter-class similarity among visually related dishes, making it a challenging fine-grained classification problem.

To ensure reproducibility, we construct a fixed 90/10 train–validation split using a deterministic random seed. This results in approximately 90,900 training images and 10,100 validation images. Prior to training, we verify data integrity by scanning all class directories and excluding any empty or invalid folders. Only valid image files are retained for training and evaluation.

For efficiency, the dataset is extracted once from a compressed TAR archive to local disk storage within the Colab runtime. This avoids repeated I/O overhead from remote storage and enables faster multi-worker data loading during training.

5.2 EXPERIMENTAL SETUP

Hardware and Runtime Environment All experiments are conducted using Google Colab with access to an NVIDIA A100 GPU, which supports bfloat16 (BF16) mixed-precision arithmetic. Mixed precision is enabled throughout training to reduce memory usage and improve computational throughput while maintaining numerical stability.

Training Configuration Training is performed for a maximum of 20 epochs, with early stopping applied when validation accuracy does not improve for three consecutive epochs. We adopt a two-stage resolution strategy: an initial training phase at 192×192 resolution, followed by an optional fine-tuning phase at 224×224 resolution to recover fine-grained spatial details.

Key training hyperparameters include:

- Batch size of 128 for 192×192 training and 96 for 224×224 fine-tuning
- AdamW optimizer with layer-wise learning rates
- Four data-loading workers with pinned memory on local disk
- BF16 mixed-precision training using automatic mixed precision (AMP)
- Gradient clipping with an ℓ_2 norm threshold of 1.0

Layer-wise Learning Rates To stabilize fine-tuning and prevent catastrophic forgetting, different learning rates are applied to different network components:

- Classifier head: 3×10^{-4}
- Residual block `layer4`: 1×10^{-4}
- Residual block `layer3`: 5×10^{-5}
- Residual block `layer2`: 2×10^{-5}

A cosine annealing learning-rate schedule is used to smoothly decay learning rates over training.

5.3 EVALUATION METRICS

Model performance is evaluated using **top-1 classification accuracy** on the held-out validation set, which measures the proportion of samples for which the predicted class matches the ground-truth label.

In addition, we monitor training and validation loss curves to assess convergence behavior and detect overfitting. Although Food-101 is sometimes evaluated using top-5 accuracy, this work focuses primarily on top-1 accuracy, which better reflects real-world food recognition scenarios.

Table 1: Validation accuracy for different fine-tuning strategies on Food-101.

Method	**Top-1 Accuracy (%)**
Head only (fc)	∼65
Layer4 + head	∼71
Layer2–4 + head	∼76
Layer2–4 + head + MixUp	∼77
+ 224 fine-tuning	**78.5**

5.4 Baselines and Ablation Studies

To quantify the contribution of each design choice, we compare the proposed method against several progressively stronger baselines. All baselines use the same ResNet-18 backbone pretrained on ImageNet.

Baseline Models The following baselines are evaluated:

- **Head-only fine-tuning**: All backbone layers are frozen and only the classifier head is trained.
- **Layer4 + head**: Only the final residual block and classifier head are unfrozen.
- **Layer2–4 + head**: Residual blocks `layer2`, `layer3`, and `layer4` are unfrozen along with the classifier head.
- **Layer2–4 + head + MixUp**: MixUp augmentation is applied during training to improve generalization.
- **Layer2–4 + head + MixUp + 224 fine-tuning**: A short high-resolution fine-tuning stage is added after convergence at 192×192.

Results Overview Table 1 summarizes validation accuracy for each baseline. Performance improves consistently as additional backbone layers are unfrozen and regularization techniques are introduced. The optional 224×224 fine-tuning stage yields the highest validation accuracy of 78.5%, demonstrating the benefit of recovering fine-grained spatial information.

6 Results

We evaluate the proposed approach on the Food-101 validation split using top-1 classification accuracy. This section presents quantitative performance, training dynamics during the final fine-tuning stage, and qualitative error analysis to better understand model behavior.

6.1 Validation Accuracy

The final model achieves a top-1 validation accuracy of **78.5%** after a short high-resolution fine-tuning phase at 224×224. This represents a substantial improvement over the baseline fine-tuning configuration trained at lower resolution, confirming the effectiveness of deeper backbone adaptation and high-resolution refinement. Similar benefits of late-stage resolution scaling have been observed in prior visual recognition studies He et al. (2016); Touvron et al. (2021).

6.2 Training Dynamics at High Resolution

Figure 1 shows the training and validation accuracy curves during the 224×224 fine-tuning phase. Unlike the earlier 192×192 stage, this phase is intentionally short and uses a reduced learning rate to refine high-frequency visual details without overfitting.

Although training accuracy remains relatively low due to strong regularization (MixUp Zhang et al. (2018) and label smoothing Szegedy et al. (2016)), validation accuracy consistently improves across epochs. This behavior indicates that the model prioritizes generalization over memorization, which is desirable in fine-grained visual recognition tasks Goodfellow et al. (2016).

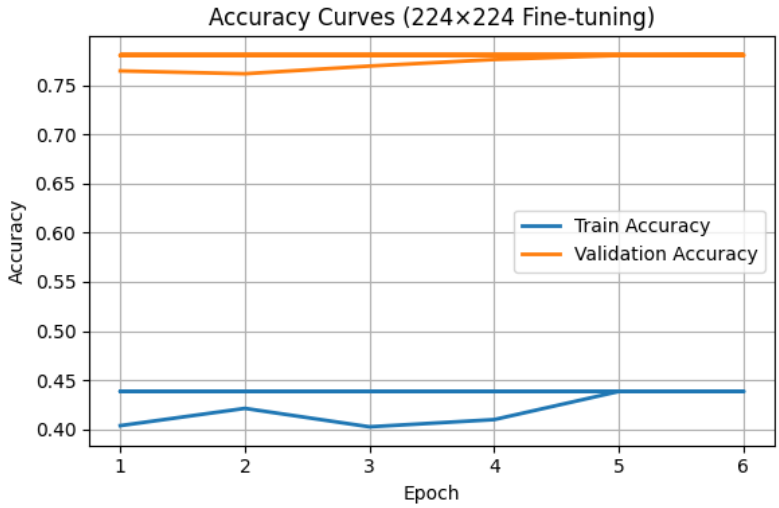

Figure 1: Training and validation accuracy versus epoch during the 224×224 fine-tuning phase.

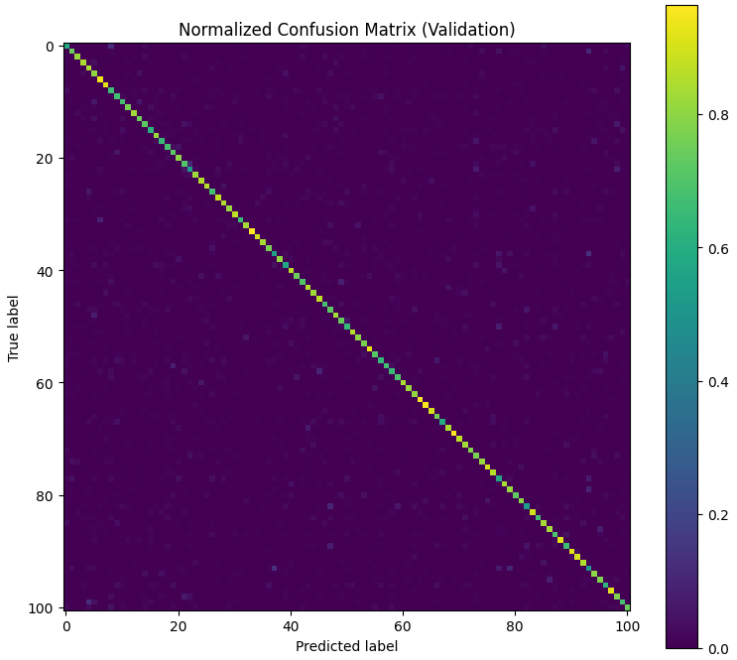

Figure 2: Normalized confusion matrix for the top-10 most frequent misclassified class pairs.

6.3 Confusion Analysis

To analyze remaining failure cases, we compute a confusion matrix on the validation set and extract the ten most frequent misclassification pairs. Figure 2 visualizes the normalized confusion matrix restricted to these top error pairs.

The dominant confusions occur between visually similar food categories such as *filet mignon* vs. *steak*, *tuna tartare* vs. *beef tartare*, and *chocolate cake* vs. *chocolate mousse*. These errors reflect the high inter-class similarity and subtle preparation differences inherent to food recognition, as previously noted in the original Food-101 benchmark Bossard et al. (2014).

True Class	Predicted Class	Count
filet_mignon	steak	15
steak	filet_mignon	12
tuna_tartare	beef_tartare	12
chocolate_cake	chocolate_mousse	11

Table 2: Top-10 most frequent confusion pairs on the Food-101 validation set.

6.4 Qualitative Analysis

Figure 3 presents representative examples of correct predictions from the validation set. In these cases, the model successfully captures global structure and key visual cues, such as dish composition, plating style, and dominant ingredients.

Figure 3: Correct predictions on the Food-101 validation set. The true class (T) and predicted class (P) with confidence are shown.

Figure 4 highlights the most confident incorrect predictions. Many of these errors correspond to visually ambiguous dishes or overlapping culinary categories. Notably, the model often assigns high confidence to semantically related alternatives, indicating that misclassifications are rarely arbitrary but instead reflect plausible visual ambiguity.

Figure 4: Most confident incorrect predictions on the Food-101 validation set. Despite high confidence, predictions often correspond to visually similar classes.

Overall, the qualitative results align with the quantitative analysis: while the model demonstrates strong generalization and robustness, the remaining errors predominantly arise from fine-grained inter-class similarity rather than systematic failure modes.

6.5 Summary of Findings

The results demonstrate that combining selective backbone unfreezing, strong regularization, and a short high-resolution fine-tuning phase significantly improves performance. Despite using a compact architecture (ResNet-18), the proposed training strategy achieves competitive accuracy while remaining computationally efficient and stable in a cloud-based training environment.

7 Discussion

This work shows that strong performance on fine-grained food image classification can be achieved using a compact architecture when paired with carefully designed fine-tuning and regularization strategies. Starting from a pretrained ResNet-18 backbone, progressive unfreezing, layer-wise optimization, and multi-resolution training yield substantial improvements over naive transfer learning baselines.

One key observation is that unfreezing intermediate and deeper residual blocks (layers 2–4) consistently improves validation accuracy compared to training only the classifier head or final block. This suggests that mid-level representations capturing texture and structural patterns are particularly important for food recognition, where inter-class visual differences are subtle Bossard et al. (2014); He et al. (2016). Layer-wise learning rates further stabilize this process by allowing higher layers to adapt while preserving lower-level features.

Regularization plays a central role in generalization. Label smoothing and MixUp both contribute measurable gains, with MixUp notably reducing overconfidence and improving robustness to am-

biguous visual cues Zhang et al. (2018); Szegedy et al. (2016). These effects are especially beneficial given the high intra-class variability and visual similarity characteristic of the Food-101 dataset.

A short high-resolution fine-tuning stage at 224×224 provides a final accuracy boost despite limited additional training. While training accuracy remains suppressed due to strong regularization, validation accuracy improves consistently, indicating better generalization rather than overfitting. This supports prior findings that late-stage resolution scaling can recover fine-grained visual details Touvron et al. (2021).

Error analysis reveals that remaining failures largely occur between visually similar categories, such as filet mignon vs. steak or tuna tartare vs. beef tartare. Qualitative inspection shows that many high-confidence errors correspond to semantically plausible alternatives, highlighting the inherent difficulty of fine-grained food classification rather than systematic model failure.

From a practical perspective, this study emphasizes training stability and efficiency in constrained environments. Mixed-precision training, checkpointing, and conservative learning-rate schedules enable reliable experimentation on cloud platforms without requiring larger architectures or excessive compute. Overall, the results suggest that thoughtful fine-tuning can significantly narrow the performance gap between compact models and more computationally intensive approaches.

Future work may explore ingredient-level supervision, self-supervised pretraining, or joint prediction of nutritional attributes, as well as attention-based or transformer architectures to further improve discrimination among visually similar dishes.

8 Limitations and Practical Considerations

Although the Food-101 dataset provides nutritional metadata, this work focuses exclusively on the image classification task. Nutritional information is used only as a post-hoc lookup after prediction and does not influence model training or optimization. Integrating nutritional estimation directly into the learning objective is left for future work.

All experiments were conducted in a cloud-based notebook environment with constrained session duration and GPU availability. Consequently, careful checkpointing, mixed-precision training, and selective layer fine-tuning were necessary to ensure training stability and reproducibility. While larger architectures or longer training schedules may further improve performance, such approaches would require more substantial computational resources. Exploring scalability to larger backbones or dedicated hardware platforms remains an important direction for future investigation.

9 Conclusion and Future Work

This work demonstrates that competitive performance on fine-grained food image classification can be achieved using a compact convolutional architecture when paired with carefully designed fine-tuning strategies. Starting from a pretrained ResNet-18 backbone, we show that progressively unfreezing intermediate residual blocks, applying strong regularization, and performing a short high-resolution refinement stage substantially improve performance over standard transfer learning baselines. Our final model achieves a top-1 validation accuracy of **78.5%** on the Food-101 dataset while remaining computationally efficient and stable within a cloud-based training environment.

Experimental results and ablation studies highlight the importance of adapting mid- and high-level feature representations for food recognition, where visually similar classes often differ only in subtle texture or preparation cues. Regularization techniques such as MixUp and label smoothing consistently improve generalization, while layer-wise learning rates stabilize optimization during deeper fine-tuning. A brief fine-tuning phase at 224×224 further recovers fine-grained visual details without inducing overfitting, underscoring the value of resolution scaling as a late-stage refinement strategy.

Despite these gains, remaining errors primarily occur between semantically related and visually similar categories, indicating that fine-grained distinctions remain challenging for standard convolutional features. Future work could explore incorporating self-supervised pretraining, lightweight attention mechanisms, or transformer-based architectures to improve discrimination in such cases. Additionally, extending the model to jointly predict food categories and nutritional attributes presents

an interesting direction for multi-task learning. Overall, this study shows that thoughtfully engineered fine-tuning and regularization strategies enable compact models to achieve strong performance on challenging visual recognition tasks under practical compute constraints.

REFERENCES

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 446–461, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Google Research. Google colaboratory. https://colab.research.google.com, 2023. Accessed via cloud-hosted Jupyter notebooks with GPU support.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1097–1105, 2012.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2019.

Paulius Micikevicius et al. Mixed precision training. In *ICLR*, 2018.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.

Hugo Touvron et al. Training data-efficient image transformers & distillation through attention. *International Conference on Machine Learning (ICML)*, 2021.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

Zhun Zhong et al. Random erasing data augmentation. In *AAAI*, 2020.

A APPENDIX

This appendix provides additional analyses and visualizations that complement the main results presented in the paper. These figures offer deeper insight into class-level performance, model confidence behavior, and qualitative predictions, but are omitted from the main body for clarity and space constraints.

A.1 PER-CLASS ACCURACY ANALYSIS

Figure 5 shows the best- and worst-performing food categories based on per-class validation accuracy. While certain visually distinctive classes such as *hot dog* and *lasagna* achieve high accuracy, other classes with strong visual overlap (e.g., *filet mignon*, *bread pudding*) remain challenging. This highlights the impact of fine-grained inter-class similarity in food recognition.

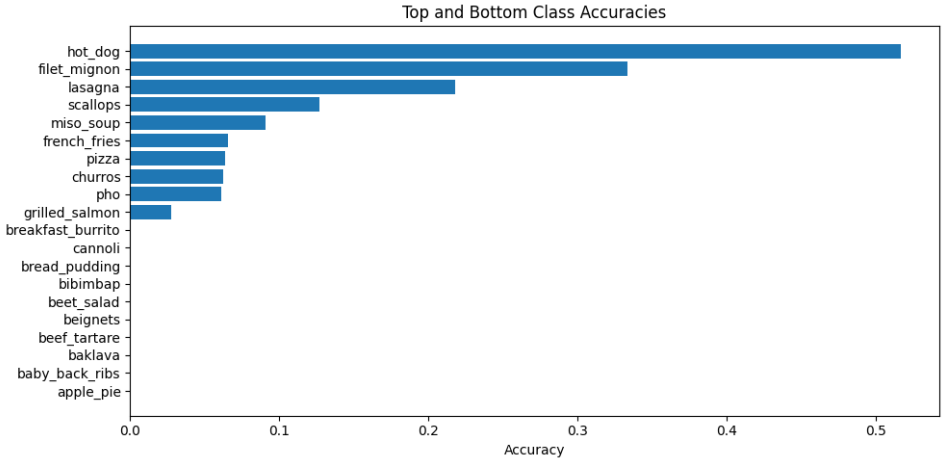

Figure 5: Top and bottom Food-101 classes ranked by per-class validation accuracy.

A.2 Prediction Confidence Distribution

To assess model calibration and confidence behavior, Figure 6 presents a histogram of prediction confidence scores on the validation set. The distribution indicates that most predictions fall within a moderate confidence range, reflecting the effect of strong regularization techniques such as MixUp and label smoothing, which discourage overconfident outputs.

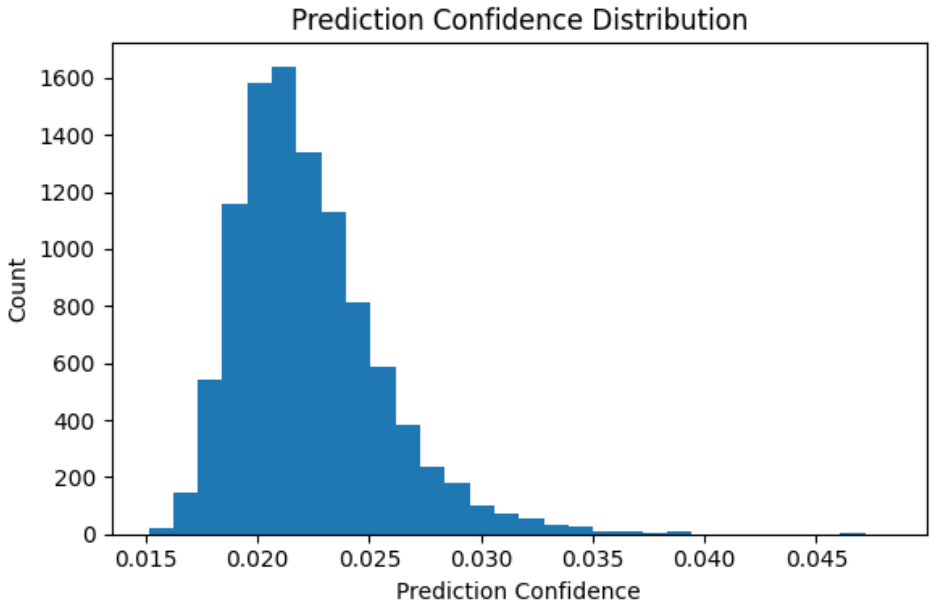

Figure 6: Distribution of prediction confidence values on the Food-101 validation set.

A.3 Additional Qualitative Results

Figure 7 provides additional qualitative examples of model predictions. The top row illustrates correct predictions across a variety of food categories, while the bottom row highlights representative misclassifications. Incorrect predictions typically correspond to visually similar or semantically related dishes, reinforcing observations from the confusion analysis in the main text.

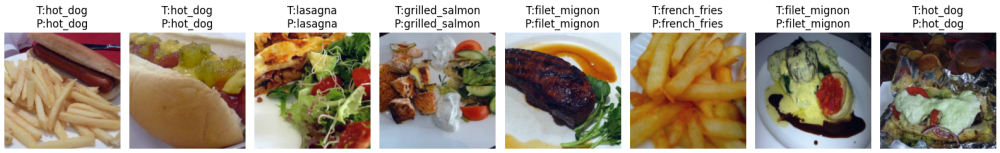

Figure 7: Additional qualitative examples from the validation set. Correct predictions (top) and incorrect predictions (bottom) are shown with true and predicted labels.

A.4 Random Validation Samples

For completeness, Figure 8 shows randomly selected images from the Food-101 validation set. These samples illustrate the diversity of visual appearance, background clutter, and presentation styles present in the dataset.

Figure 8: Randomly sampled images from the Food-101 validation set.