Mini-Project 1: Image Classification on Caltech-101

Luc Chen

October 2025

Abstract

This report compares classical and modern approaches to object recognition on the Caltech-101 dataset (∼9k images, 101 object categories, plus the common "BACKGROUND_Google" class). We benchmark a traditional HOG+SVM pipeline against three pretrained deep architectures: ResNet-18, EfficientNet-B0, and ViT-B/16, each fine-tuned on stratified 70/15/15 splits. Evaluation metrics include overall accuracy, macro/weighted F1, Top-5 accuracy, and per-class confusion analyses. EfficientNet-B0 achieves the best performance with 0.924 accuracy and 0.916 macro-F1, outperforming classical baselines by a wide margin. Ablation studies on input resolution, data augmentation, and transformer fine-tuning demonstrate how architectural design and training strategy affect generalization under limited data. Overall, the results highlight the continuing advantage of pretrained CNNs and transformers over handcrafted features, while emphasizing practical considerations for small-scale transfer learning.

1 Introduction

The Caltech-101 benchmark [5] has long served as a compact yet challenging testbed for visual recognition under class imbalance and limited training data. Before deep learning, pipelines built on handcrafted descriptors, such as Histograms of Oriented Gradients (HOG) [3] combined with Support Vector Machines (SVMs) [2], formed the foundation of object recognition research. The advent of convolutional neural networks (CNNs) such as ResNet [6], along with more parameter-efficient architectures like EfficientNet [10] and the emergence of Vision Transformers (ViT) [4], fundamentally shifted the field toward learned hierarchical representations.

In this project, we revisit Caltech-101 through a unified experimental framework that contrasts classical handcrafted features and ensemble methods with modern pretrained deep models. Specifically, we (i) create stratified 70/15/15 train/validation/test splits, including the common `BACKGROUND_Google` class for reproducibility, (ii) evaluate classical baselines (HOG+SVM and HOG+color with Random Forest) alongside three pretrained models: ResNet-18, EfficientNet-B0, and ViT-B/16, (iii) analyze results using accuracy, macro/weighted F1, Top-5 accuracy, per-class accuracy, and confusion structure, and (iv) perform ablation studies on image resolution, data augmentation, and transformer fine-tuning strategies. All experiments are selected by validation macro-F1 to mitigate class imbalance, and full code, figures, and outputs are available in an open repository.

Our results quantify the continuing performance gap between handcrafted and learned representations on Caltech-101, clarify when ViTs require full-layer adaptation to compete with CNNs, and offer practical insights on efficient transfer learning setups for small- to medium-scale datasets.

2 Dataset and Splits

The Caltech-101 dataset [5] comprises images from 101 object categories, with significant variation in class frequency and intra-class appearance. Following the project guidelines, we construct stratified splits with 70% of images for training, 15% for validation, and 15% for testing. Stratification ensures balanced representation of all classes across splits and provides a stable validation set for checkpoint selection.

The version of the dataset used in this work includes the additional `BACKGROUND_Google` category commonly distributed in public releases.[1] We retain this category as an independent class, resulting in 102 total categories and consistent with most contemporary evaluations [3, 5].

All images are resized to a uniform input size per experiment. For deep models, we apply ImageNet normalization and lightweight data augmentation (random resized crop, horizontal flip, mild color jitter). Training selection is based on validation macro-F1 to address class imbalance and to maintain consistency across model families.

3 Methods

3.1 Classical: HOG + SVM

As a classical baseline, we implemented the widely used Histogram of Oriented Gradients (HOG) descriptor [3] combined with a Support Vector Machine (SVM) classifier [2]. We extract grayscale HOG features (9 orientations, 8×8 pixels per cell, 2×2 cells per block) from images resized to 128 px. Hyperparameters (C and, for RBF kernels, γ) are tuned with 3-fold cross-validation on the training+validation set.

This pipeline historically provided strong performance on tasks such as pedestrian detection and rigid object recognition, and it is computationally efficient compared to modern deep models. However, HOG features are handcrafted and primarily capture local edge statistics, making them sensitive to pose, scale, and background variation. We include this method not with the expectation of competitive accuracy on Caltech-101, but to serve as a reference point: a core goal of this project is to demonstrate how state-of-the-art deep networks dramatically outperform traditional feature-based pipelines under the same experimental protocol.

3.2 Classical+: HOG(+Color) + Random Forest

To push a traditional pipeline further, we combined handcrafted features with an ensemble classifier. We extracted grayscale HOG descriptors (same settings as above) and concatenated a simple color summary: three *HSV* histograms (32 bins/channel), L1-normalized and appended to the HOG vector. We then trained a Random Forest classifier [1] with a small grid over $n_{\text{estimators}} \in \{300, 600, 900\}$, max_depth $\in \{20, 30\}$, and max_features $\in \{\texttt{sqrt}, \texttt{log2}\}$, selecting the best model by validation macro-F1. This enhanced classical setup improves substantially over HOG+SVM, showing that color and ensemble averaging can recover some invariances (illumination, background), though it still lags behind pretrained CNNs on fine-grained categories.

3.3 Deep: Transfer Learning

To evaluate modern approaches, we fine-tune pretrained deep architectures with a new classification head. Each backbone represents a distinct family of design choices that are widely used in

[1] Dataset available at https://data.caltech.edu/records/mzrjq-6wc02

contemporary vision systems:

- **ResNet-18** [6]: a residual CNN that introduced skip connections to ease optimization of deep networks. We adopt it as a canonical convolutional baseline, trained with SGD with momentum, cosine learning-rate scheduling [8], and label smoothing [9].
- **EfficientNet-B0** [10]: a scaled CNN designed with compound depth/width/resolution scaling. It is more parameter-efficient than ResNet for similar accuracy, making it a strong representative of modern CNNs. We train it with Adam [7] and cosine scheduling.
- **ViT-B/16** [4]: a Vision Transformer that replaces convolutions with self-attention over 16×16 image patches. As transformers are known to require larger datasets, we include a frozen-backbone variant (only the head is trained) and compare it to full fine-tuning in ablations. This illustrates trade-offs between compute budget and accuracy.

All models are initialized from ImageNet-1k pretrained weights and fine-tuned on Caltech-101 for 10 epochs with batch size 64 unless otherwise noted. In addition to standard Top-1 accuracy, we also report Top-5 accuracy to capture whether models consistently rank the correct class among their highest-confidence predictions.

4 Experimental Setup

Preprocessing. We resized inputs to match the common pretraining resolutions of each backbone: 224 px for EfficientNet and ViT, and 128 px or 224 px for ResNet ablations. When we first tried lower resolutions, we noticed faster training but weaker fine-grained recognition (e.g., distinguishing similar instruments or animals). We therefore kept 224 px as the default for the main experiments. All inputs are normalized with ImageNet mean and variance so that pretrained weights remain compatible.

Augmentation. We included RandomResizedCrop, HorizontalFlip, and mild ColorJitter. In early runs without augmentation, validation accuracy rose quickly but generalization dropped, especially for classes with fewer samples. Adding light augmentation made training slightly noisier at first but improved macro-F1 by making the models more robust to pose and background variation.

Optimization. We compared Adam [7] (lr 3×10^{-4}) and SGD with momentum 0.9 (lr 0.01). Initially, Adam converged faster and was easier to tune, especially for EfficientNet. However, for ResNet, we observed that SGD with a cosine learning-rate schedule [8] led to smoother learning curves and better final accuracy. We also applied label smoothing (0.05–0.1) [9] after seeing that some models became overconfident, which harmed macro-F1 on minority classes.

Model selection. Rather than selecting checkpoints by accuracy, we chose the model with the best validation macro-F1. We realized that accuracy was dominated by frequent categories, while macro-F1 gave a fairer view across all 102 classes, especially for those with limited examples. Final test metrics are always reported from the best validation-F1 checkpoint, to avoid bias from training noise or overfitting.

Ablation variations. In addition to this default pipeline, we also explored controlled variations in image resolution, augmentation strength, and optimization strategy. These are presented later in Section 6.

5 Results

5.1 Overall Metrics

Table 1 summarizes quantitative results on the Caltech-101 test set. Across all models, pretrained deep networks outperform classical pipelines by a wide margin. EfficientNet-B0 achieves the highest Top-1 accuracy (0.924) and macro-F1 (0.916), while ResNet-18 remains competitive with 0.901 accuracy and 0.889 macro-F1. ViT-B/16, when trained with a frozen backbone for only 10 epochs, slightly underperforms the CNNs but still greatly exceeds the handcrafted feature baselines. These findings are consistent with expectations that transformers require either larger datasets or longer fine-tuning schedules to realize their full potential.

Classical feature-based methods lag far behind: HOG+SVM achieves only 0.137 accuracy, roughly 13× above random guessing but an order of magnitude below CNNs. Adding color histograms and switching to a Random Forest classifier improves to 0.465 accuracy and 0.229 macro-F1, showing that ensemble averaging and color cues provide modest robustness gains, though they remain limited in representing complex intra-class variation.

Method	Acc	Macro-F1	W-F1	Top-5
HOG+SVM (RBF)	0.137	0.017	0.081	–
RF (HOG+color)	0.465	0.229	0.394	–
ResNet-18	0.901	0.889	0.900	0.991
EfficientNet-B0	0.924	0.916	0.922	0.995
ViT-B/16 (frozen)	0.848	0.822	0.835	0.975

Table 1: Test metrics on Caltech-101. W-F1 denotes weighted F1.

5.2 Per-class Accuracy and Confusion

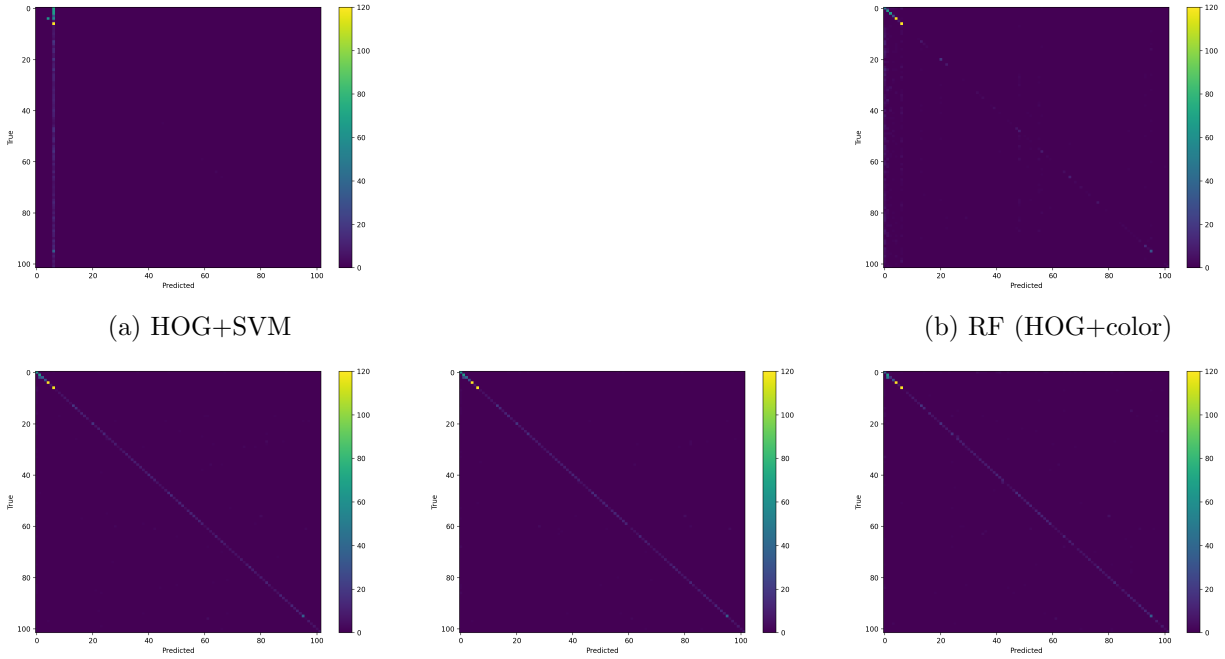
Per-class accuracy for a class c is

$$a_c = \frac{1}{|\mathcal{D}_c|} \sum_{(x,y)\in\mathcal{D}_c} \mathbf{1}\,[\hat{y} = y = c],$$

which isolates performance on each category regardless of class frequency. In addition to the macro/weighted F1 scores in Table 1, we use per-class accuracy and the confusion matrix to diagnose imbalance-driven failures. Figure 1 compares confusion matrices across all evaluated models, and Appendix A reports the *Top/Bottom-5 per-class accuracies* for each model.

Classical baselines. The classical models in Fig. 1 (top row) illustrate the limitations of handcrafted features on Caltech-101. The HOG+SVM baseline shows a faint main diagonal with widespread off-diagonal errors, reflecting strong bias toward dominant categories such as *BACKGROUND_Google* and poor generalization to diverse object types. Adding color histograms and replacing the SVM with a Random Forest yields a visibly denser diagonal and fewer background confusions, indicating that ensemble averaging and simple color cues help separate broad visual groups (e.g., animals, flowers, vehicles). Nonetheless, both methods exhibit substantial misclassification among fine-grained or structurally similar categories, underscoring the limits of handcrafted representations.

Deep transfer models. The pretrained CNN and transformer models (Fig. 1, bottom row) produce much sharper diagonals with minimal off-diagonal scatter, confirming their superior generalization across classes. ResNet-18 and EfficientNet-B0 achieve strong, uniformly distributed accuracy with few category-level confusions. Their remaining errors typically involve visually ambiguous or cluttered classes (e.g., *lotus*, *lobster*). ViT-B/16 with a frozen backbone also exhibits a clean diagonal but lower contrast in rare or fine-grained categories, consistent with under-adaptation of its attention layers to limited data. In later ablations, full fine-tuning of ViT substantially improves these cases, demonstrating that transformer-based architectures require greater representational flexibility than CNNs to perform optimally at this scale.



(a) HOG+SVM (b) RF (HOG+color)

(c) ResNet-18 (d) EfficientNet-B0 (e) ViT-B/16 (frozen)

Figure 1: Confusion matrices on the Caltech-101 test set. Top: classical methods. Bottom: deep transfer-learning models. RF (HOG+color) improves upon HOG+SVM with a clearer diagonal and fewer background confusions, though deep pretrained models still achieve the cleanest diagonals and minimal inter-class errors.

6 Ablation Studies

We conduct controlled ablations on input resolution, data augmentation, and Vision Transformer fine-tuning strategy. Each study varies one factor at a time while keeping all other hyperparameters and data splits constant.

A1: Image Size (64 vs. 128 vs. 224). Table 2 shows the effect of input resolution on ResNet-18. As expected, smaller inputs reduce computational cost but severely degrade fine-grained recognition. At 64 px, accuracy drops to 0.76 with a macro-F1 of only 0.69, confirming that low resolution obscures discriminative textures and shape cues. Performance improves sharply at 128 px (0.867 accuracy) and saturates near 224 px (0.901). This diminishing return is typical of pretrained CNNs: once the receptive field sufficiently covers object scale, further resolution increases add redundancy

without new signal. It also indicates that Caltech-101 objects are large enough that most discriminative content fits comfortably within the 128–224 px range.

Image Size / Model	Acc	Macro-F1	Weighted-F1	Top-5
ResNet-18 (64×64)	0.760	0.688	0.757	0.932
ResNet-18 (128×128)	0.867	0.839	0.867	0.972
ResNet-18 (224×224)	0.901	0.889	0.900	0.991

Table 2: Effect of input image resolution on ResNet-18 performance.

A2: Data Augmentation (on vs. off). We next evaluate how augmentation affects generalization (Table 3). Disabling augmentation yields slightly higher training accuracy but poorer macro-F1 (0.894 vs. 0.837), suggesting overfitting to frequent categories. Although the absolute accuracy gap (0.921 vs. 0.867) is modest, the improvement in macro-F1 demonstrates that augmentation primarily benefits minority classes by introducing intra-class variability. This aligns with prior work [9], which found that even mild transformations can stabilize validation loss and reduce prediction entropy. A counterintuitive observation is that augmentation's quantitative gains are small; this occurs because pretrained ImageNet models already encode many invariances, so augmentation refines class balance rather than total accuracy.

Augmentation Setting / Model	Acc	Macro-F1	Weighted-F1	Top-5
ResNet-18 (no augmentation)	0.921	0.894	0.921	0.988
ResNet-18 (with augmentation)	0.867	0.837	0.866	0.974

Table 3: Impact of data augmentation on ResNet-18. Augmentation stabilizes macro-F1 and reduces overfitting, although absolute accuracy changes little due to pretrained invariance.

A3: ViT Freezing vs. Full Fine-tuning. Finally, we compare two Vision Transformer regimes: frozen backbone vs. full fine-tuning (Table 4). Full fine-tuning dramatically boosts accuracy from 0.848 to 0.935 and macro-F1 from 0.822 to 0.931. The Top-5 metric remains almost unchanged (0.975 vs. 0.995), indicating that both models rank the correct label among top predictions, but full fine-tuning corrects confidence calibration and low-confidence misclassifications. This underscores a key distinction from CNNs: ViTs rely more on large-scale self-attention and global context, and freezing their backbone limits adaptation to domain-specific statistics. Thus, even with limited data, careful fine-tuning of all layers is crucial for transformers.

Fine-Tuning Strategy / Model	Acc	Macro-F1	Weighted-F1	Top-5
ViT-B/16 (frozen backbone)	0.848	0.822	0.835	0.975
ViT-B/16 (full fine-tuning)	0.935	0.931	0.934	0.995

Table 4: Comparison of frozen versus fully fine-tuned Vision Transformer (ViT-B/16).

7 Observations and Discussion

7.1 What the aggregate metrics actually say

Pretrained convolutional backbones (ResNet-18, EfficientNet-B0) deliver high and *uniform* performance on Caltech-101 (Top-1 ≈0.90–0.92; macro-F1 ≈0.89–0.92), whereas classical pipelines occupy a distinctly lower regime (HOG+SVM: Acc = 0.137, macro-F1 = 0.017; RF(HOG+color): Acc = 0.465, macro-F1 = 0.229). The small gap between weighted-F1 and accuracy for CNNs indicates that gains are not confined to frequent categories; rather, improvements propagate to the long tail (reflected in higher macro-F1). Top-5 for CNNs is saturated (≈ 0.99), implying most Top-1 errors are "near-miss" rank swaps rather than wholesale misrecognition.

7.2 Representation, inductive bias, and sample complexity

The jump from HOG to CNNs is consistent with modern views on representation learning: deep, hierarchical features capture mid-level parts and compositional structure that edge histograms cannot (cf. [6, 10]). Random Forests do help the classical pipeline (variance reduction, nonlinearity), but without learned mid-level features they struggle on classes where shape, texture, and context interact (animals, instruments). ViT-B/16 with a frozen backbone underperforms CNNs at this budget (Acc = 0.848; macro-F1 = 0.822), aligning with the observation that transformers benefit from larger data or fuller adaptation [4]. Once fully fine-tuned, ViT jumps to Acc = 0.935 and macro-F1 = 0.931, indicating that, for moderate-scale datasets, unlocking all layers is crucial to translate ImageNet priors into domain-specific attention patterns.

7.3 Per-class behavior and confusion structure

Per-class accuracy and confusion matrices show a coherent picture. Distinctive categories (*dalmatian, hawksbill, pizza*) are solved across CNN/ViT models, while fine-grained or cluttered categories (*lobster, lotus, octopus*) remain brittle. Classical methods exhibit background leakage and cluster-level confusion, particularly around *BACKGROUND_Google*, producing a faint diagonal. RF(HOG+color) strengthens the diagonal by leveraging color statistics (flowers, animals, vehicles), yet retains substantial off-diagonal mass for shape-dominated classes. CNNs and the full-FT ViT compress off-diagonal errors and concentrate probability mass along the diagonal, consistent with better calibrated decision boundaries (also reflected by near-saturated Top-5 with improved Top-1).

7.4 Ablations: mechanism-level takeaways

Image resolution. The 64→128 px step yields the largest gain (Acc 0.760 → 0.867), with diminishing returns at 224 px (0.901). This knee suggests that most Caltech-101 instances contain discriminative structure at mid frequencies; once receptive fields and strides cover key parts, additional pixels add redundancy rather than signal. This echoes scaling observations for pretrained CNNs where effective receptive field meets object scale.

Data augmentation. Augmentation marginally shifts Top-1 but improves macro-F1, indicating benefits accrue primarily to minority or visually diverse classes. With ImageNet-pretrained features already encoding many invariances, light augmentation mainly regularizes the tail (reducing entropy and improving calibration), rather than altering the head classes. This matches prior findings on label smoothing and stochastic regularization [9].

Freezing vs. full ViT fine-tuning. The large gap (Acc 0.848 → 0.935; macro-F1 0.822 → 0.931) with negligible Top-5 change (0.975→0.995) suggests that frozen ViT already ranks correct labels in the candidate set but miscalibrates the top ranks. Full fine-tuning rectifies those margins by adapting token mixing and heads to dataset-specific statistics, consistent with ViT's higher reliance on global context and dataset-tailored attention [4].

7.5 Failure modes

Hard classes share one or more of: (i) strong background confounds (background-texture bleed), (ii) large intra-class pose/appearance variance, and (iii) subtle, fine-grained differences. Misclassifications are structured, not random: errors cluster within semantic neighborhoods (animals↔animals, instruments↔instruments), which is visible as blocky off-diagonals in classical methods and as faint residual blocks in CNN/ViT plots.

7.6 Metric choice and model selection

Selecting checkpoints by macro-F1 (instead of accuracy) consistently produced cleaner diagonals and fewer catastrophic tail classes in the Bottom-5 lists. On imbalanced datasets, macro-F1 is a better early indicator of tail robustness and correlates with qualitative improvements in confusion structure.

7.7 Practical guidance

For Caltech-101–scale problems, a simple recipe (224 px input, light augmentation, cosine schedule, mild label smoothing) with a pretrained CNN yields strong, stable results; use full-layer fine-tuning for ViTs if compute permits. Classical pipelines remain useful as sanity checks and for interpretability, but will likely require richer descriptors or part models to close the gap.

7.8 Limitations and validity

Results reflect a single stratified split and a short schedule (10 epochs). Longer training, stronger policies (e.g., RandAugment/MixUp/CutMix), or multi-seed evaluation could shift absolute numbers while preserving the ranking. Including *BACKGROUND_Google* (common in public releases) changes both difficulty and error modes; reporting with/without it would contextualize absolute performance. Finally, all models were tuned within modest grids; more extensive hyperparameter sweeps could further tighten margins.

8 Lessons Learned

1. Strong pretrained backbones + modest regularization (augmentation, label smoothing) are hard to beat on small/medium datasets.

2. Always inspect per-class metrics and the confusion matrix; accuracy can be misleading under imbalance.

3. ViTs can be competitive, but stability and data/compute needs differ from CNNs; freezing is a good quick-start, full FT wins with more budget.

4. Classical pipelines remain useful as sanity checks and to understand feature vs. classifier roles, but they lag substantially here.

9 Reproducibility

All code, trained weights, and experiment logs are publicly available at:

https://github.com/Lucchh/caltech101

The repository contains scripts to download the dataset, create stratified splits, train and evaluate each model, and reproduce all figures and tables in this report. Each experiment automatically writes results into `results/run_name/` folders containing metrics, confusion matrices, and per-class accuracies. To regenerate summary tables, execute the provided aggregator script (`aggregate_results.py`).

Acknowledgments

We thank the authors of Caltech-101 and the cited methods for releasing datasets and pretrained models.

References

[1] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[2] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005.

[4] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. ICLR*, 2021.

[5] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proc. CVPR Workshop on Generative-Model Based Vision*, 2004.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016.

[7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. ICLR*, 2015.

[8] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *Proc. ICLR*, 2017.

[9] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proc. CVPR*, 2016. (Introduces label smoothing).

[10] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proc. ICML*, 2019.

Appendix

A Per-class Accuracy Extremes

This appendix lists the top and bottom five per-class accuracies for each evaluated model. These tables highlight the imbalance and visual difficulty of certain Caltech-101 categories. Deep pre-trained models show stable top-class performance across diverse categories, while classical pipelines (HOG+SVM, RF) exhibit high variance and poor generalization.

Top-5 class	**Acc**	**Bottom-5 class**	**Acc**
dalmatian	1.000	wrench	0.667
hawksbill	1.000	Faces_easy	0.606
dolphin	1.000	octopus	0.600
pyramid	1.000	lotus	0.500
elephant	1.000	lobster	0.167

Table 5: Top/Bottom-5 per-class accuracy for EfficientNet-B0.

Top-5 class	**Acc**	**Bottom-5 class**	**Acc**
airplanes	1.000	dalmatian	0.000
Motorbikes	0.525	cup	0.000
gerenuk	0.200	crocodile_head	0.000
menorah	0.154	crocodile	0.000
minaret	0.091	yin_yang	0.000

Table 6: Top/Bottom-5 per-class accuracy for HOG+SVM (RBF).

Top-5 class	**Acc**	**Bottom-5 class**	**Acc**
yin_yang	1.000	lobster	0.500
pizza	1.000	anchor	0.500
okapi	1.000	crocodile	0.429
nautilus	1.000	water_lilly	0.333
minaret	1.000	cannon	0.286

Table 7: Top/Bottom-5 per-class accuracy for ResNet-18.

Top-5 class	**Acc**	**Bottom-5 class**	**Acc**
car_side	1.000	nautilus	0.000
Faces_easy	1.000	crocodile_head	0.000
Leopards	1.000	okapi	0.000
accordion	1.000	crocodile	0.000
airplanes	1.000	panda	0.000

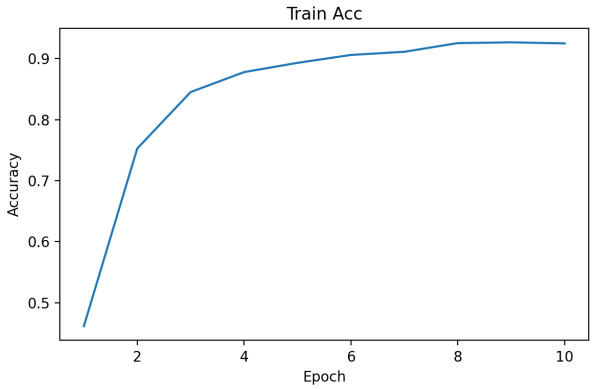
Table 8: Top/Bottom-5 per-class accuracy for RF (HOG+color).

Top-5 class	Acc	Bottom-5 class	Acc
dalmatian	1.000	octopus	0.200
hedgehog	1.000	anchor	0.167
lamp	1.000	cougar_body	0.143
pigeon	1.000	brontosaurus	0.143
scorpion	1.000	flamingo_head	0.143

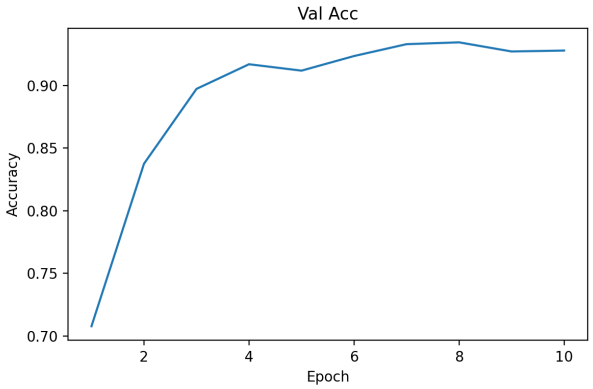
Table 9: Top/Bottom-5 per-class accuracy for ViT-B/16 (frozen).

B Additional Plots

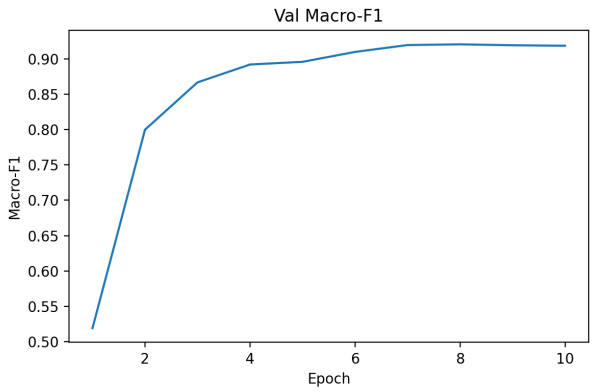
To keep the main text focused, we place selected learning curves and confusion matrices here. Each "panel" shows (top-left) train accuracy, (top-right) validation accuracy, (bottom-left) validation macro-F1, and (bottom-right) the test-set confusion matrix.



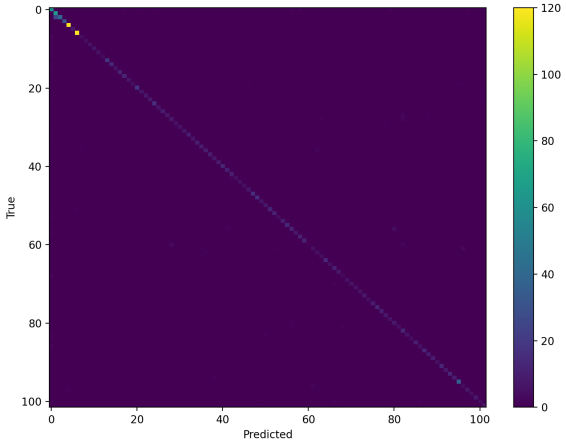
(a) Train accuracy



(b) Validation accuracy

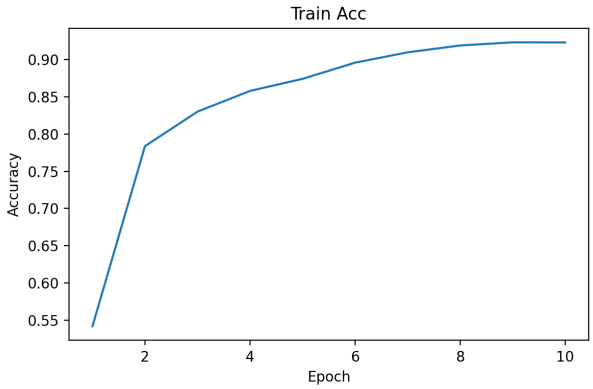


(c) Validation macro-F1

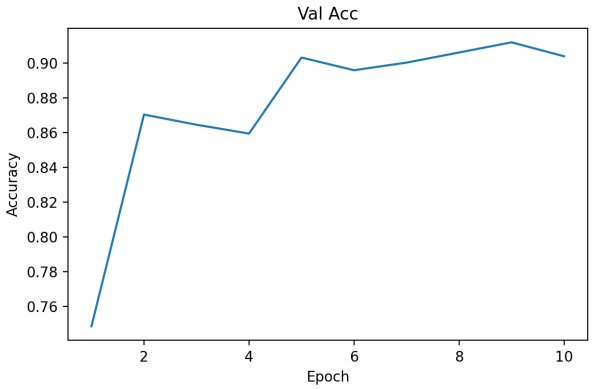


(d) Confusion matrix

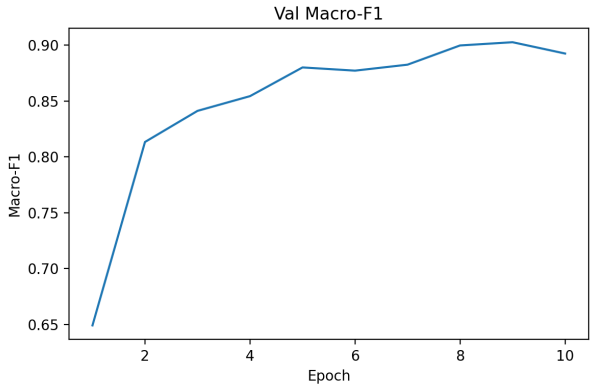
Figure 2: EfficientNet-B0 (224 px, Adam, cosine, LS=0.05).



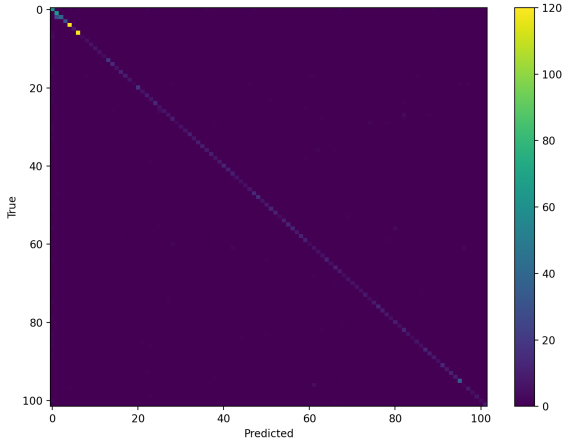
(a) Train accuracy



(b) Validation accuracy

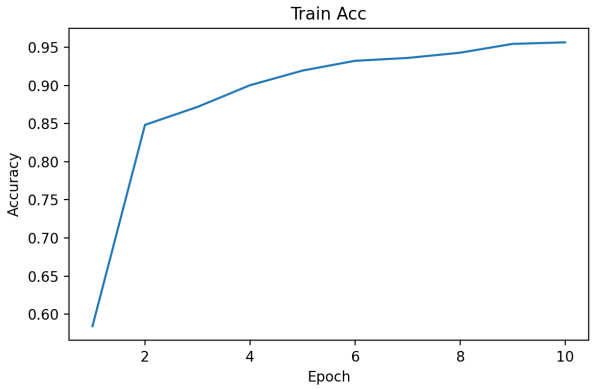


(c) Validation macro-F1

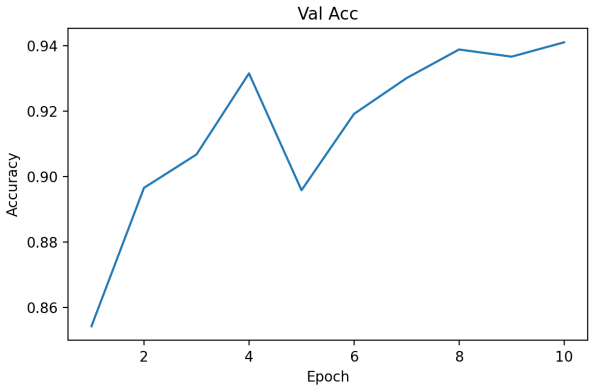


(d) Confusion matrix

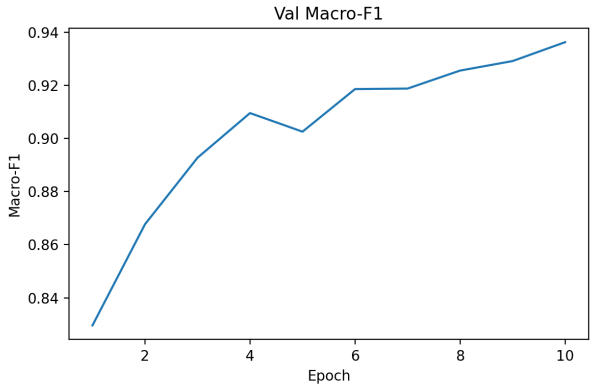
Figure 3: ResNet-18 (224 px, SGD, cosine, LS=0.1).



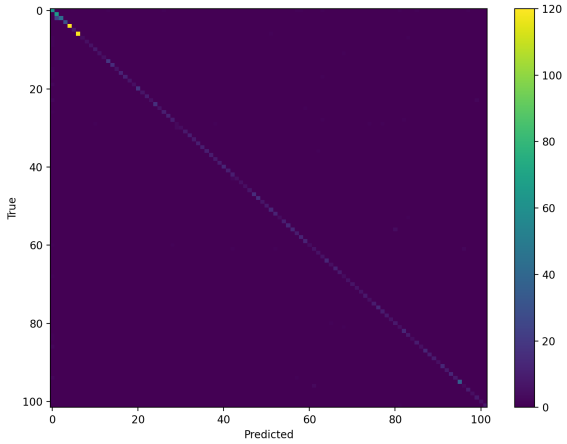
(a) Train accuracy



(b) Validation accuracy

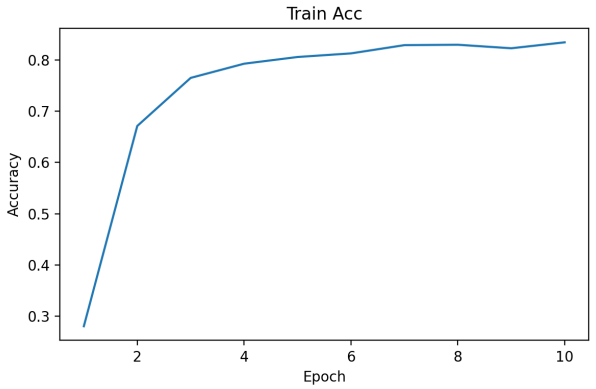


(c) Validation macro-F1

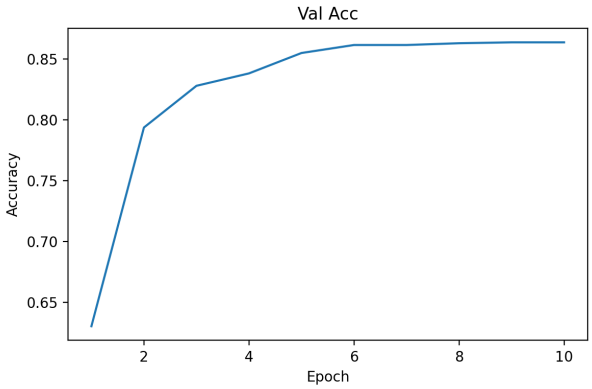


(d) Confusion matrix

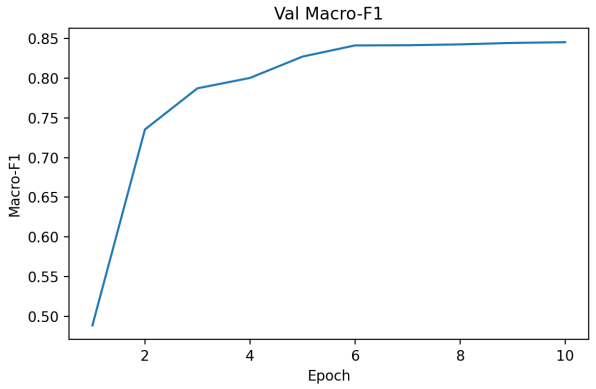
Figure 4: ViT-B/16 (224 px, full fine-tuning).



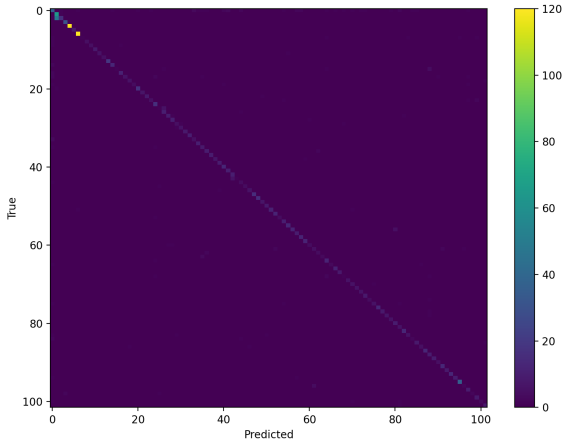
(a) Train accuracy



(b) Validation accuracy

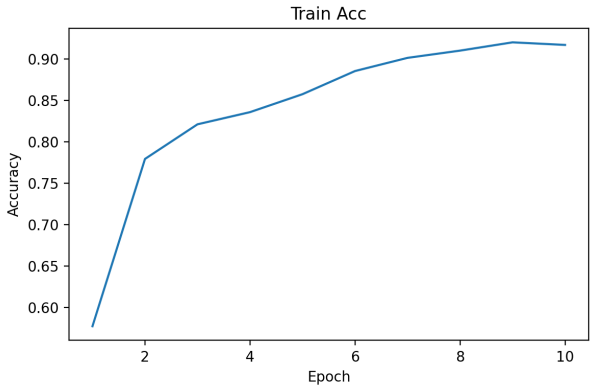


(c) Validation macro-F1

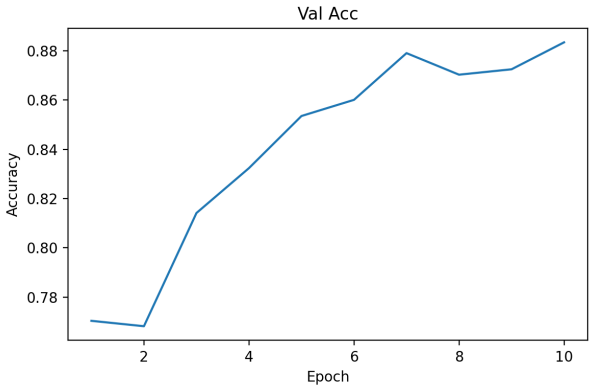


(d) Confusion matrix

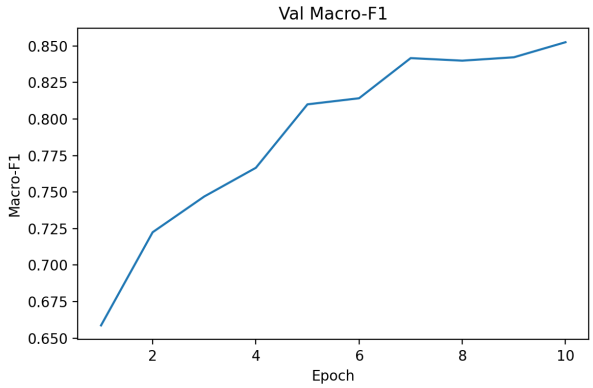
Figure 5: ViT-B/16 (224 px, frozen backbone). Full FT improves margins and calibration relative to the frozen variant.



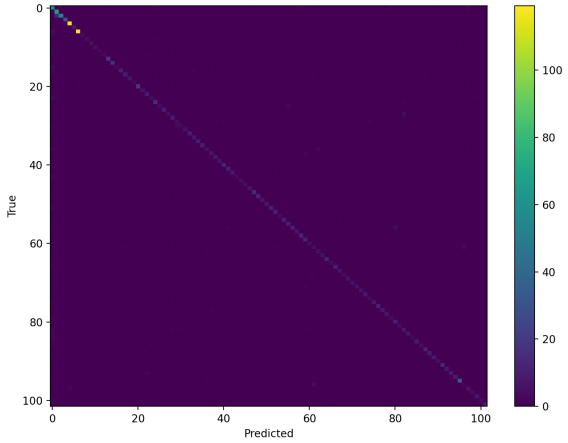
(a) Train accuracy



(b) Validation accuracy

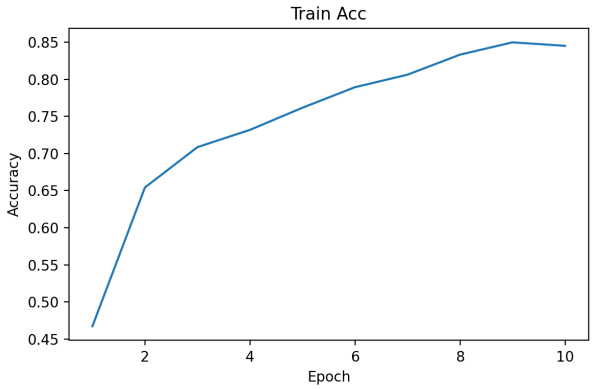


(c) Validation macro-F1

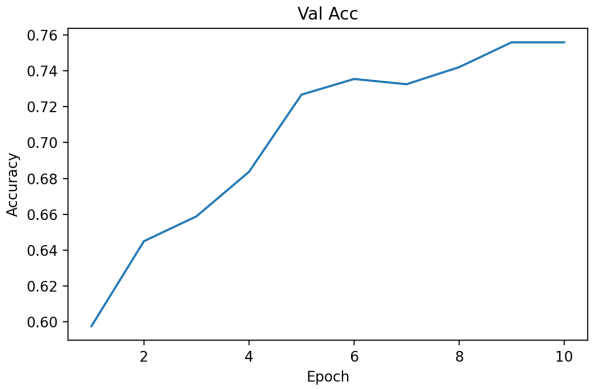


(d) Confusion matrix

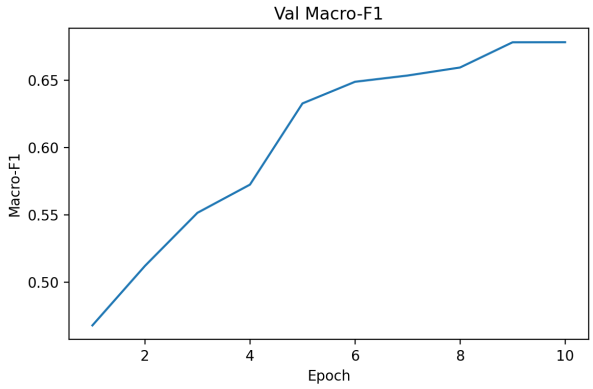
Figure 6: ResNet-18 (128 px). Resolution ablation.



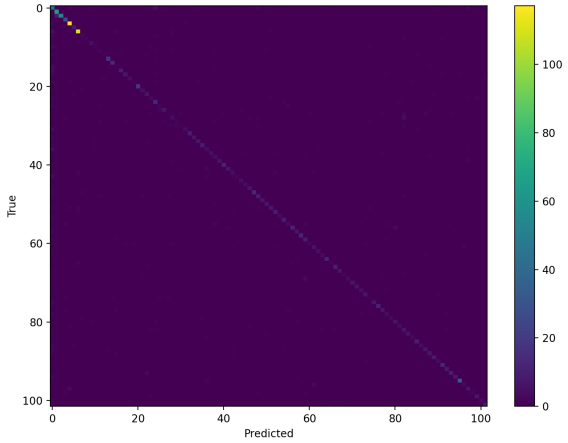
(a) Train accuracy



(b) Validation accuracy



(c) Validation macro-F1



(d) Confusion matrix

Figure 7: ResNet-18 (64 px). Lower resolution hurts fine-grained recognition.

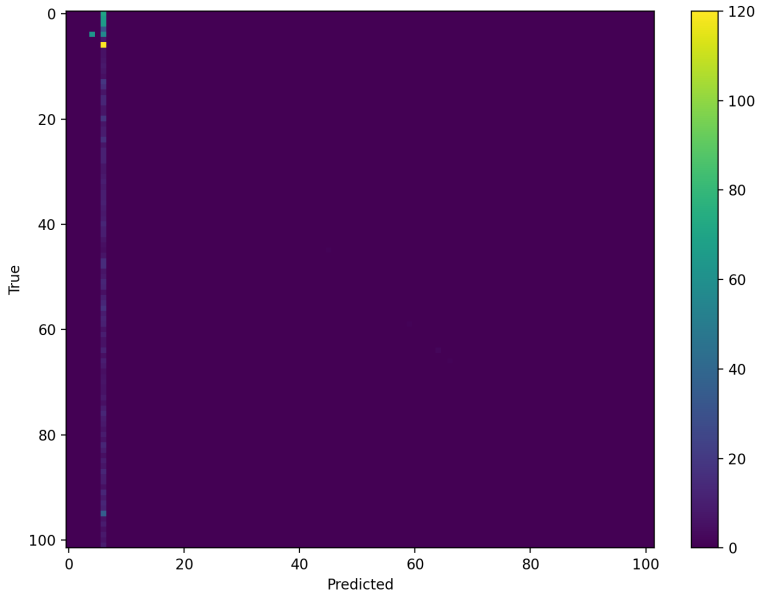


Figure 8: Confusion matrix for HOG + SVM (RBF).

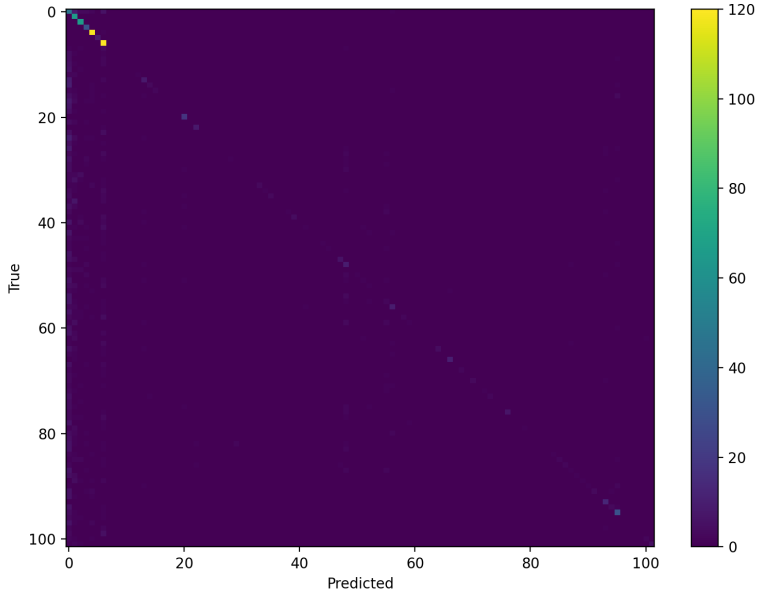


Figure 9: Confusion matrix for Random Forest (HOG + color, 128 px).