

Caltech-101 Classification Report: Classical Baseline, Transfer Learning, and Ablation Studies

March 1, 2026

Abstract

This report summarizes training and evaluation results saved in `runs_server` for Caltech-101 classification over 102 classes (including `BACKGROUND_Google`). We compare a classical baseline (HOG + Linear SVM) against transfer-learning models (ResNet50, EfficientNet-B0, ViT-B/16), then analyze ablations on image size, augmentation, and optimizer. On the main test runs, ViT-B/16 achieves the best Top-1 accuracy (0.9461) and Macro-F1 (0.9272), followed by ResNet50 (0.9402, 0.9183) and EfficientNet-B0 (0.9213, 0.8955). The classical SVM baseline remains much lower (Top-1 0.5036), but provides a lightweight reference point. Ablations show strong dependence on image resolution for CNNs, limited gains from the tested augmentation setup, and a large optimizer sensitivity for ResNet50 (Adam \gg SGD in this setup).

1 Introduction

Caltech-101 is an important benchmark for multi-class object recognition, and it has been widely used to evaluate both classical and modern visual representation learning methods [1]. Traditional pipelines often combine hand-crafted Histogram of Oriented Gradients (HOG) descriptors [2] with Support Vector Machine (SVM) classifiers [3], while recent progress is dominated by deep architectures such as ResNet [4], EfficientNet [5], and Vision Transformer (ViT) [6].

In this work, we evaluate Caltech-101 classification under a unified experimental framework using all 102 classes, including `BACKGROUND_Google`, with a stratified split protocol (train/val/test = 6400/1372/1372 images, i.e., 70.0%/15.0%/15.0%). We implement a classical baseline (HOG + Linear SVM) and multiple transfer-learning deep models (ResNet50, EfficientNet-B0, and ViT-B/16), report Top-1 accuracy, Top-5 accuracy, Macro-F1, and Weighted-F1, and run ablation studies on image resolution, augmentation policy, and optimizer choice.

The purpose of this report is to provide a rigorous empirical comparison between classical and deep-learning approaches on the same split protocol, identify practical factors that most affect performance, and summarize actionable insights for model selection on medium-scale, class-imbalanced object recognition benchmarks.

2 Dataset

All experiments use Caltech-101 [1], with 9,144 images across 102 classes. We construct data splits with a two-stage stratified procedure over class labels: first, 70% training and 30% temporary data; second, the temporary set is split evenly into validation and test sets, yielding 6,400/1,372/1,372 images (70.0%/15.0%/15.0%). This preserves class proportions across splits and avoids sample overlap.

We intentionally keep `BACKGROUND_Google` as a valid class, rather than discarding it, to better reflect realistic recognition conditions where images may not correspond to a clean foreground object category. Including this class makes the task more challenging and more repre-

sentative, and it allows us to evaluate whether models can separate true object categories from heterogeneous background-like content.

Preprocessing is model-specific but standardized within each family. For the classical pipeline, each image is first converted to RGB to standardize mixed source formats into a consistent 3-channel input, then resized to 128×128 to keep feature dimensionality fixed, scaled to $[0, 1]$ for numerical stability, converted to grayscale so HOG focuses on intensity-gradient structure, and finally encoded with HOG features (9 orientations, 8×8 pixels per cell, 2×2 cells per block). For deep models, ResNet50 and EfficientNet-B0 are trained at 128×128 , while ViT-B/16 is run at 224×224 to satisfy architecture input requirements; all deep inputs are normalized with ImageNet mean/std to match pretrained feature statistics, and training uses either deterministic resize without augmentation or RandomResizedCrop + horizontal flip + color jitter with augmentation for regularization, while validation/test always use deterministic resize and normalization for consistent evaluation.

3 Methods

3.1 HOG + Linear SVM

As a classical baseline, we employ a linear Support Vector Machine (SVM) on HOG representations. Model selection is performed on the validation split by maximizing Macro-F1 over the regularization parameter C , and the selected model is evaluated on the held-out test set.

3.2 HOG + Engineered Features + LightGBM

To provide a stronger non-linear classical comparator, we concatenate HOG features with nine engineered image statistics (width, height, aspect ratio, channel-wise RGB means, and channel-wise RGB standard deviations) and train a multiclass LightGBM classifier. Hyperparameters (`num_leaves`, `learning_rate`, `n_estimators`) are selected using validation Macro-F1.

3.3 Transfer-Learning Deep Models

We fine-tune three ImageNet-pretrained backbones and adapt each architecture to 102-way classification by replacing the final prediction head.

3.3.1 ResNet50

ResNet50 is a residual convolutional architecture in which skip connections enable stable optimization of deep networks. We replace the final fully connected layer with a 102-class classifier and fine-tune the model with cross-entropy, selecting the checkpoint that achieves the best validation Macro-F1.

3.3.2 EfficientNet-B0

EfficientNet-B0 is a compact convolutional model that balances depth, width, and resolution through compound scaling. We replace its classifier layer with a 102-class head and fine-tune it under the same training/selection criterion, retaining the best validation Macro-F1 checkpoint.

3.3.3 ViT-B/16

ViT-B/16 is a transformer-based vision architecture that models global image context via self-attention over patch tokens. We replace the transformer head with a 102-class classifier and fine-tune the model with cross-entropy, using the best validation Macro-F1 checkpoint for test evaluation.

3.4 Training Protocol

For all deep models, training lasts 12 epochs with a two-stage freeze–unfreeze strategy: the backbone is frozen for the first 2 epochs and then unfrozen for full fine-tuning. ResNet50 and EfficientNet-B0 use batch size 32, initial learning rate 3×10^{-4} , and post-unfreeze learning rate 1×10^{-4} at 128×128 resolution, whereas ViT-B/16 uses batch size 16, initial learning rate 1×10^{-4} , and post-unfreeze learning rate 5×10^{-5} at 224×224 ; in all deep runs, Adam is the default optimizer with weight decay 1×10^{-4} .

For the classical models, HOG extraction uses size 128, 9 orientations, 8×8 pixels per cell, and 2×2 cells per block. The linear SVM is trained with `max_iter=1000` and validation tuning over C (configured grid: [0.1]), while the LightGBM model searches over `num_leaves` $\in \{31, 63\}$, `learning_rate` $\in \{0.05, 0.1\}$, and `n_estimators` $\in \{200, 400\}$ with `subsample=0.9` and `colsample_bytree=0.9`.

4 Experiment Setup

4.1 Augmentation

Two settings are tested in A2:

- **Without augmentation:** deterministic resize pipeline
- **With augmentation:** random resized crop + horizontal flip + color jitter

4.2 Optimization

Adam is used as the default optimizer. A3 compares Adam vs SGD for ResNet50 under the same image size and augmentation setting.

4.3 Model Selection

Deep model comparison uses the main runs under `runs_server/default_exp`. Classical baseline is reported from `classical_svm` in the same directory for consistency.

4.4 Ablation Protocol

- **A1 (Image Size):** compare 64 vs 128 (ViT remains effectively at 224 by architecture constraint).
- **A2 (Augmentation):** compare without augmentation vs with augmentation at fixed image size.
- **A3 (Optimizer):** compare Adam vs SGD (ResNet50).

5 Results

5.1 Overall Performance

Table 1: Main test results from `runs_server/default_exp`.

Model	Top-1 Acc	Top-5 Acc	Macro-F1	Weighted-F1
Classical SVM	0.5036	0.6334	0.3043	0.4741
EfficientNet-B0	0.9213	0.9869	0.8955	0.9205
ResNet50	0.9402	0.9942	0.9183	0.9402
ViT-B/16	0.9461	0.9913	0.9272	0.9459

5.2 Per-Class Results

Average per-class accuracy across the 102 classes:

- Classical SVM: 0.2893
- EfficientNet-B0: 0.8977
- ResNet50: 0.9222
- ViT-B/16: 0.9350

Additionally, 37 classes are predicted perfectly (accuracy = 1.0) by all three deep models.

Table 2: ViT-B/16 hardest classes on test set (lowest per-class accuracy).

Class	Accuracy	Test Count
lotus	0.3000	10
crocodile	0.5714	7
snoopy	0.6000	5
wrench	0.6667	6
BACKGROUND_Google	0.7000	70

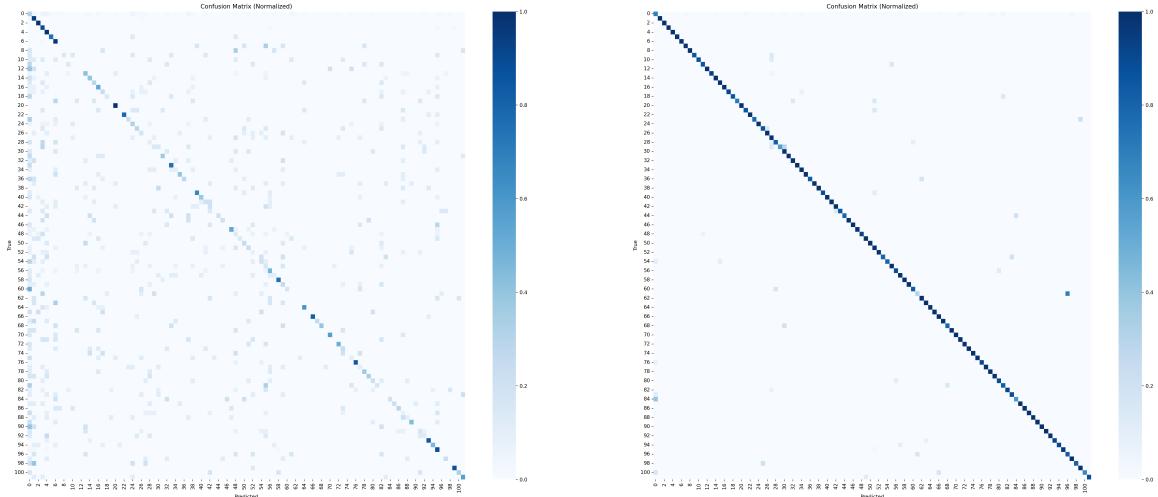


Figure 1: Confusion matrix comparison between classical baseline and best deep model.

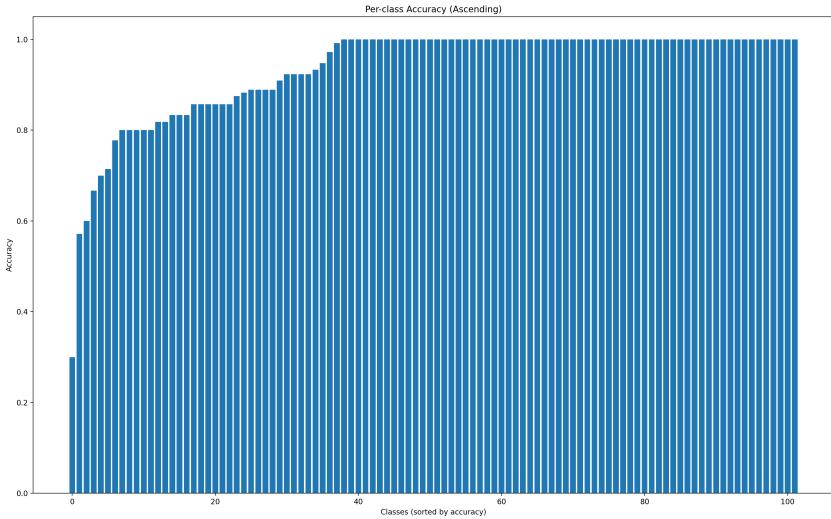


Figure 2: Per-class accuracy distribution for ViT-B/16.

6 Ablation Study

6.1 A1: Image Size (64 vs 128)

Table 3: A1 image-size ablation (Top-1 accuracy).

Model	Acc@64	Acc@128	$\Delta(128-64)$
ResNet50	0.8739	0.9402	+0.0663
EfficientNet-B0	0.8061	0.9213	+0.1152
ViT-B/16	0.9461	0.9461	+0.0000

6.2 A2: Augmentation (Without vs With)

Table 4: A2 augmentation ablation (Top-1 accuracy).

Model	Without Aug	With Aug	$\Delta(\text{With}-\text{Without})$
ResNet50	0.9410	0.9402	-0.0007
EfficientNet-B0	0.9359	0.9213	-0.0146
ViT-B/16	0.9592	0.9461	-0.0131

6.3 A3: Optimizer (Adam vs SGD on ResNet50)

Table 5: A3 optimizer ablation on ResNet50 (Top-1 accuracy).

Optimizer	Top-1 Acc	Δ vs Adam
Adam	0.9402	0.0000
SGD	0.8411	-0.0991

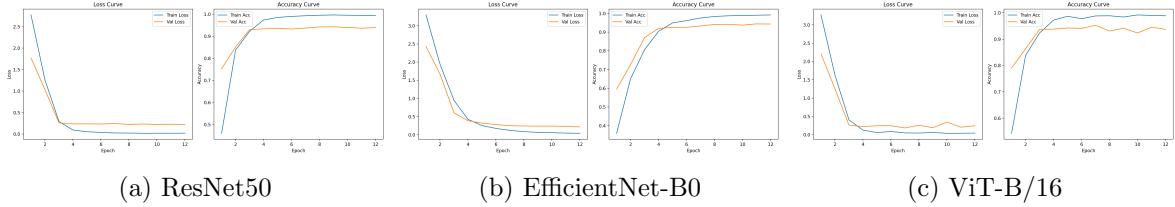


Figure 3: Training curves for main deep-learning runs.

7 Observation and Discussion

- **Deep models strongly outperform classical baseline.** The gap between ViT (0.9461) and SVM (0.5036) is large in both Top-1 and Macro-F1, showing the benefit of pretrained deep representations for this 102-class setting.
- **ViT is the best model overall.** ViT achieves the highest Top-1 and Macro-F1, and the best mean per-class accuracy.
- **Image resolution matters for CNNs.** Increasing from 64 to 128 yields notable gains for ResNet50 and EfficientNet-B0, especially EfficientNet (+0.1152).
- **Current augmentation policy is not beneficial.** In this setup, adding augmentation slightly decreases Top-1 across all three deep models, suggesting that augmentation strength/type should be retuned.
- **Optimizer choice is critical.** For ResNet50 under the tested hyperparameters, Adam clearly outperforms SGD by nearly 10 percentage points Top-1.
- **Per-class variance remains visible.** Even with strong overall performance, classes such as `lotus`, `crocodile`, and `snoopy` remain difficult, likely due to limited samples and/or higher intra-class variability.

References

- [1] 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 *CVPR Workshop on Generative-Model Based Vision*, 2004.
- [2] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.
- [3] C. Cortes and V. Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.
- [4] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [5] M. Tan and Q. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.
- [6] A. Dosovitskiy, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*, 2021.