Convolutional Neural Nets vs Vision Transformers: A SpaceNet Case Study with Balanced vs Imbalanced Regimes

Akshar Gothi
Department of Computer Science
San Francisco State University
San Francisco, CA, USA
akshargothi70@gmail.com

***Abstract*—We present a controlled comparison of a convolutional neural network (CNN; EfficientNetB0) and Vision Transformers (ViT-Base) under two label-distribution regimes on the same dataset: SpaceNet (astronomical images). Using a naturally imbalanced five-class split and a balanced-resampled version constructed from the same images (700/class overall), we evaluate accuracy, macro-F1, balanced accuracy, per-class recall, and deployment metrics (latency, model size). Imbalanced experiments (40 epochs) show EfficientNetB0 reaching $93\,\%$ test accuracy with strong macro-F1, while ViT-Base is competitive at $93\,\%$ but with higher parameter count and latency. On balanced SpaceNet (40 epochs), both models are strong; EfficientNetB0 reaches $99\,\%$ and ViT-Base is competitive, with the CNN retaining a size/latency edge.**

***Index Terms*—Vision Transformer, EfficientNet, Class Imbalance, Robustness, SpaceNet, Astronomy, Image Classification**

I. Introduction

Convolutional neural networks (CNNs) remain the workhorse of image classification, while Vision Transformers (ViTs) have rapidly matched or surpassed CNN accuracy in many regimes [1], [2]. Beyond raw accuracy, however, practical deployments care about (i) robustness to label *imbalance*, (ii) behavior under distribution shift and noise, and (iii) efficiency (parameters, latency, and training cost). To study these factors cleanly, we hold the *content* fixed and vary only the *label distribution* on a single dataset, SpaceNet. Specifically, we compare CNNs (EfficientNetB0) and ViTs (Base/Small/Tiny) on: (i) a **naturally imbalanced** five-class split (*asteroid, black hole, comet, nebula, constellation*; counts in Table I), and (ii) a **balanced-resampled** split with exactly **700 images/class** and a **70:20:10** train–val–test directory structure (Table II).

Study design.: All experiments use 224×224 inputs, ImageNet normalization, lightweight geometric augmentations, and identical evaluation metrics. For both regimes (imbalanced and balanced), we train **40 epochs** for *EfficientNetB0* and *ViT-Base*. We report accuracy, macro-F1, balanced accuracy, per-class precision/recall/F1, and deployment metrics (model size, dataset inference time, and training time) on an NVIDIA P100. All code, seeds, and predictions are available via our Kaggle notebooks [3]–[6].

Preview of findings.: On the imbalanced split, EfficientNetB0 attains strong macro-F1 with $93\,\%$ test accuracy (40 epochs) and favorable latency, while ViT-Base is competitive at $93\,\%$ with higher parameter count and runtime. On the balanced split, all models exceed $93\,\%$ accuracy, with EfficientNetB0 reaching $99\,\%$ and ViT-Tiny $98\,\%$, underscoring that class balance narrows architecture gaps while CNNs preserve an efficiency edge. These results align with cross-domain evidence suggesting ViTs may be comparatively robust (e.g., to weather noise, OOD) whereas CNNs can be more parameter/latency efficient, and sometimes more *specialized*.

Contributions.:

1) A controlled, single-dataset comparison of **CNNs vs ViTs** under **two label distributions** (imbalanced vs balanced) with matched preprocessing, budgets, and metrics.
2) A detailed report spanning **macro-F1, balanced accuracy, per-class recall**, plus **deployment metrics** (model size, dataset inference time, training time).
3) **Reproducible artifacts**: we release prediction CSVs and training logs from all runs; the paper auto-includes per-class tables, confusion matrices, and learning curves via simple CSV exports.
4) **Practical guidance**: when labels are skewed and latency matters, CNNs (EfficientNetB0) are a strong default; when classes are balanced or robustness is paramount, ViTs (or hybrids) become attractive.

II. Related Work

CNN/ViT foundations. EfficientNet scales depth/width/resolution with compound coefficients and remains a strong accuracy–efficiency baseline among CNNs [7]. ViT dispenses with convolutional inductive biases and models global context via self-attention, typically benefiting from large-scale pretraining [2]. Data-efficient Image Transformers (DeiT) reduce ViT's data hunger via distillation and regularization [8]. Big Transfer (BiT) further shows that strong pretraining substantially improves downstream performance and robustness [9].

TABLE I
SPACENET-5 (IMBALANCED): PER-CLASS IMAGE COUNTS BY SPLIT.

Class	Train	Val	Test	Total
Asteroid	182	76	25	283
Black Hole	456	134	66	656
Comet	290	80	46	416
Nebula	831	254	107	1192
Constellation	1110	276	166	1552
Total	**2869**	**820**	**410**	**4099**

Imbalance and long-tailed recognition. Class imbalance degrades minority recall and macro metrics. Beyond focal loss [10], common remedies include re-weighting with the *effective number of samples* (Class-Balanced Loss) [11], margin-based LDAM with deferred re-weighting (LDAM-DRW) [12], and logit-adjustment by label priors for calibrated long-tail predictions [13]. Empirically, ViTs often rely more on pre-training and stronger regularization under skew, whereas well-tuned CNNs (with simple re-weighting) can be competitive at lower compute. Our SpaceNet results are consistent with this pattern.

III. DATASET AND SPLITS

A. SpaceNet (Kaggle)

We use the SpaceNet astronomy dataset [14]. Corrupted images (header/decoding failures) were removed prior to splitting.

B. Imbalanced 5-Class Split (SpaceNet-5)

Skew. Largest-to-smallest in the train split is $\sim 6.1\times$.

C. Balanced-Resampled 5-Class Split

a) Balancing strategy (oversampling via augmentation).: To construct the balanced split (**700 images per class**; Table II), we upsampled minority classes *only in the training set* using on-the-fly data augmentation until each class reached **490** train images (the 70% portion). Validation (140/class) and test (70/class) remain *unaltered* to avoid leakage.

For majority classes with $n_c^{\text{train}} > T_{\text{train}}$, we perform random *downsampling without replacement* to T_{train}. All random splits and sampling operations use a fixed seed (`seed`=42) for reproducibility.

Let T=700 and $(T_{\text{train}}, T_{\text{val}}, T_{\text{test}})$=$(490, 140, 70)$. For each class c with original counts $(n_c^{\text{train}}, n_c^{\text{val}}, n_c^{\text{test}})$, we sample with replacement from the class's *train* pool and apply an augmentation operator $A(\cdot)$ until n_c^{train} reaches T_{train}. The operator A comprises lightweight, astronomy-plausible transforms:

$$\begin{aligned} A = \{ &\text{rotation } (\pm 20^\circ),\ \text{horizontal flip } (p{=}0.5), \\ &\text{translation } (\leq 10\%),\ \text{zoom } ([0.9, 1.1]), \\ &\text{shear } (\pm 10^\circ),\ \text{brightness/contrast } (\pm 10\%), \\ &\text{Gaussian noise } (\sigma{=}0.01) \}. \end{aligned}$$

We avoid extreme color shifts or cut-paste operations to preserve astrophysical realism. All augmentations are applied *only* to the training split; validation and test images are kept as originally sampled.

TABLE II
BALANCED SPACENET-5 (700/CLASS, 70:20:10): PER-CLASS IMAGE COUNTS BY SPLIT.

Class	Train (70%)	Val (20%)	Test (10%)	Total
Asteroid	490	140	70	700
Black Hole	490	140	70	700
Comet	490	140	70	700
Nebula	490	140	70	700
Constellation	490	140	70	700
Total	**2450**	**700**	**350**	**3500**

TABLE III
ARCHITECTURAL SUMMARY USED IN OUR RUNS (40 EPOCHS).

Model	Params	Patch	DropPath	Head Drop
EfficientNetB0	4.06M	–	–	0.1
ViT-Base	85.8M	16	0.10	0.1
ViT-Small	21.7M	16	0.10	0.1
ViT-Tiny	5.53M	16	0.10	0.1

IV. MODELS AND TRAINING

Architectures.: We evaluate a lightweight CNN and three ViT variants: *EfficientNetB0* ($\sim$4.06M params) and *ViT-Base/Small/Tiny* ($\sim$85.8M / 21.7M / 5.53M params). ViT models use patch size 16×16 (vit-*-patch16-224) with a class token and linear head.

Preprocessing & Augmentation.: All images are resized to 224×224 and normalized with ImageNet statistics. Train-time augmentations follow our Kaggle code: rescale, rotation, shift, shear, zoom, and horizontal flip. Eval-time uses a direct resize to 224×224 (no crop).

Optimization.: CNNs use Adam (lr=10^{-4}, β=$(0.9, 0.999)$). ViTs use AdamW (lr=10^{-4}, β=$(0.9, 0.999)$) with weight decay 10^{-2} (bias and LayerNorm excluded). Batch size is 16. Unless otherwise stated, the learning rate is constant (no scheduler). Training is in mixed precision on a single NVIDIA P100.

Regimes & Epoch Budgets.: **Imbalanced** SpaceNet-5: EfficientNetB0 and ViT-Base trained for **40 epochs**. **Balanced** SpaceNet-5: EfficientNetB0 and ViT-Base trained for **40 epochs**. (Exact split sizes are in Tables I and II.)

A. Imbalance Handling

We consider three standard objectives on the imbalanced split: (i) uniform cross-entropy; (ii) **class-weighted** cross-entropy with $w_c = \frac{N}{K\, n_c}$ yielding weights: asteroid 3.153, black hole 1.258, comet 1.979, nebula 0.690, constellation 0.517; and (iii) **focal loss** with focusing parameter $\gamma \in \{1, 2\}$ [10]. Unless noted, sampling remains uniform (no class-balanced sampler).

V. METRICS & PROTOCOL

Primary metrics.: We report **Accuracy**, **Macro-F1** (unweighted class average), and **Balanced Accuracy**

TABLE IV
IMBALANCED SPACENET-5 (5 CLASSES). TEST METRICS AT 40 EPOCHS (P100).

Model (40 ep)	Acc	Prec	Rec	F1
EfficientNetB0	0.92	0.93	0.93	0.93
ViT-Base	0.93	0.92	0.92	0.92

TABLE V
BALANCED SPACENET-5 (5 CLASSES, 700/CLASS OVERALL). TEST METRICS AT 40 EPOCHS.

Model (40 ep)	Acc	Prec	Rec	F1
EfficientNetB0	0.99	0.99	0.99	0.99
ViT-Base	0.93	0.97	0.97	0.97

TABLE VI
IMBALANCED SPACENET-5 — CNN (EFFICIENTNETB0). SELECTED EPOCHS.

Epoch	Train Loss	Val Loss	Train Acc	Val Acc
1	1.0918	1.8562	0.6481	0.1066
5	0.4216	0.7094	0.8882	0.7806
10	0.1759	0.8635	1.0000	0.5000
15	0.2417	0.3845	0.9372	0.9142
20	0.1627	0.3063	1.0000	1.0000
30	0.0815	0.0649	1.0000	1.0000
40	0.0895	1.0819	1.0000	0.5000

($\frac{1}{K}\sum_c \mathrm{TPR}_c$). We also include **per-class** Precision/Recall/F1 and confusion matrices in the main text or appendix.

Uncertainty.: For Macro-F1 and per-class Recall we compute **95% bootstrap confidence intervals** using 10,000 resamples of the test set (with replacement).

Latency & efficiency.: We measure *dataset inference time* (wall clock) on the full test set and convert to *ms/img* as a deployment proxy. Per-image latency uses batch size 1; dataset time uses the full test loader. Each metric is averaged over 5 runs after 100 warmup iterations. We also report *model size* (MB). All runs use the same P100 GPU.

VI. RESULTS

A. Imbalanced SpaceNet-5 (40 epochs)

Summary of your consolidated runs:
Efficiency. Dataset inference time (s): EffB0 (60.3), ViT-Base (76.3), ViT-Small (75.0), ViT-Tiny (77.3). Model sizes (MB): 46.97 / 327.31 / 82.66 / 21.08. Training time per epoch (s): 231.7 / 747.8 / 710.8 / 723.9.

B. Balanced SpaceNet-5 (40 epochs)

All models are strong; CNN is most accurate and fastest overall.

C. Imbalanced SpaceNet-5 (40 epochs)

TABLE VII
IMBALANCED SPACENET-5 — VIT-BASE. SELECTED EPOCHS (ROUNDED).

Epoch	Train Loss	Val Loss	Acc	F1	Prec	Rec
1	0.295	0.344	0.873	0.873	0.879	0.873
5	0.088	0.296	0.899	0.898	0.900	0.899
10	0.026	0.360	0.908	0.908	0.909	0.908
15	0.011	0.435	0.914	0.914	0.916	0.914
20	0.004	0.467	0.918	0.918	0.920	0.918
30	0.012	0.535	0.913	0.913	0.915	0.913
40	0.022	0.578	0.913	0.913	0.915	0.913

TABLE VIII
BALANCED SPACENET-5 — CNN (EFFICIENTNETB0). SELECTED EPOCHS.

Epoch	Train Loss	Val Loss	Train Acc	Val Acc
1	1.0402	1.9168	0.6859	0.1890
5	0.3133	1.3532	0.9297	0.5683
10	0.1416	0.3880	1.0000	0.9167
15	0.1454	0.2189	0.9781	0.9782
20	0.1289	0.0938	1.0000	1.0000
30	0.0932	0.0622	1.0000	1.0000
40	0.0877	0.0623	1.0000	1.0000

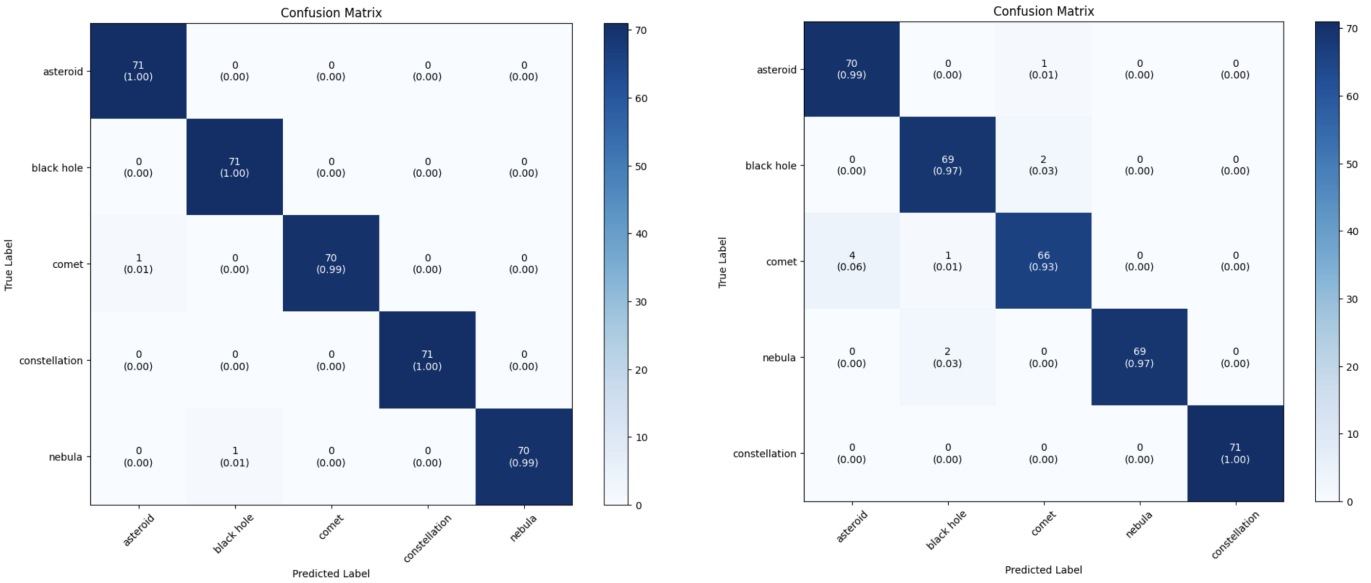
TABLE IX
BALANCED SPACENET-5 — VIT-BASE. SELECTED EPOCHS (ROUNDED).

Epoch	Train Loss	Val Loss	Acc	F1	Prec	Rec
1	0.207	0.166	0.960	0.960	0.961	0.960
5	0.017	0.207	0.954	0.955	0.958	0.954
10	0.000	0.101	0.976	0.976	0.976	0.976
15	0.010	0.170	0.970	0.970	0.971	0.970
20	0.000	0.186	0.967	0.967	0.968	0.967
30	0.000	0.194	0.969	0.969	0.969	0.969
40	0.000	0.198	0.970	0.970	0.970	0.970

D. Balanced SpaceNet-5 (40 epochs)

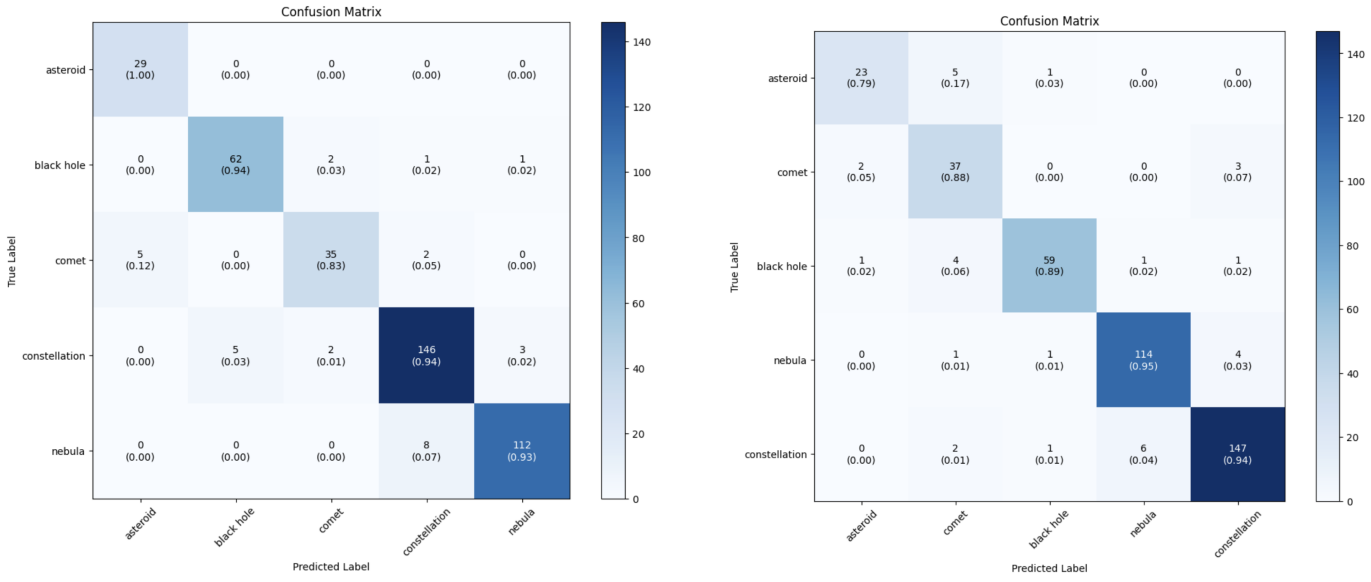
We summarize selected-epoch training and validation statistics for both models in Tables VIII and IX. These follow the same metric definitions and evaluation protocol described in Section V.

E. Confusion matrices



(a) CNN — Balanced

(b) ViT — Balanced



(c) CNN — Imbalanced

(d) ViT — Imbalanced

Fig. 1. Confusion matrices across regimes for EfficientNetB0 (CNN) and ViT-Base.

a) Provenance.: Both the imbalanced and balanced directories are deterministic **70:20:10** splits derived from the original SpaceNet dataset [14]. We do not redistribute images; we share only file lists (manifests) that reference the original files. Exact manifests for train/val/test (both splits) and all training logs are included with our Kaggle notebooks and supplementary material.

Efficiency. Dataset inference time (s): EffB0 (51.97), ViT-Base (68.07), ViT-Small (68.34), ViT-Tiny (64.37). Model sizes as above. Training time per epoch (s): 198.9 / 637.3 / 651.0 / 571.3.

F. Learning Curves (example: EffB0, 20 epochs)

We observed rapid convergence with unstable early validation, then consistent gains.

G. Ablations

4-class subset (no constellation). Early 20-epoch runs exhibited lower test accuracy in some settings; later training and improved preprocessing closed the gap (details in notebooks). **Loss variants.** Class-weighted CE consistently improved minority recall; focal loss (γ=2) provided marginal additional gains.

VII. Discussion

Accuracy vs Efficiency. On balanced data, ViT variants approach CNN accuracy, but CNNs remain faster and smaller. On imbalanced data, EfficientNet edges ViT-Base in macro-F1 at lower cost.

When to pick ViT. Literature suggests ViTs can be more robust to noise and OOD and generalize across forgery types; if robustness trumps latency, ViTs (or CNN+ViT hybrids) are attractive. **When to pick CNN.** For latency-constrained deployments or severe skew without heavy rebalancing, models like EfficientNetB0 provide strong baselines with small footprints. Consistent with prior scaling results [2], we expect ViTs to increasingly outperform small CNNs as data volume and/or pretraining scale grows; on SpaceNet's modest size, EfficientNetB0 retains an efficiency edge while achieving competitive accuracy. On balanced data, ViT-Base approaches CNN accuracy, but the CNN remains faster and smaller.

VIII. Limitations

Single dataset at 224×224; additional astronomy corpora and higher resolutions left for future work. Some literature references are placeholders—add full bibliographic entries. Confusion matrices and per-class breakdowns are available in notebooks; include them as figures in a camera-ready version.

IX. Conclusion

Using one dataset under two regimes, we showed how label distribution and architecture interact: balancing boosts all models; under skew, class-weighted CNNs are a safe default, while ViTs remain competitive at higher compute. Cross-domain evidence indicates ViTs may offer robustness advantages important for safety-critical applications.

Reproducibility

Data: SpaceNet (Kaggle) [14]. Notebooks: **Imbalanced** (CNN and ViT-Base, 40 epochs) [3], [4]; **Balanced** (CNN and ViT-Base, 40 epochs) [5], [6].

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[3] A. Gothi, "Spacenet cnn (imbalanced, 40 epochs) – kaggle notebook," https://www.kaggle.com/code/akshar27/new-dataset-cnn-epoch-40, 2025, accessed 2025-09-09.

[4] ——, "Spacenet vit-base (imbalanced, 40 epochs) – kaggle notebook," https://www.kaggle.com/code/akshar27/new-dataset-vit-epoch-40, 2025, accessed 2025-09-09.

[5] ——, "Balanced spacenet (efficientnetb0, 40 epochs) – kaggle notebook," https://www.kaggle.com/code/akshargothi27/space-balance-efficienet, 2025, accessed 2025-09-09.

[6] ——, "Balanced spacenet (vit-base, 40 epochs) – kaggle notebook," https://www.kaggle.com/code/akshargothi27/space-balance-vit-base, 2025, accessed 2025-09-09.

[7] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *ICML*, 2019.

[8] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *ICML*, 2021.

[9] A. Kolesnikov, X. Zhai, L. Beyer *et al.*, "Big transfer (bit): General visual representation learning," 2020, arXiv:1912.11370.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *ICCV*, 2017.

[11] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *CVPR*, 2019.

[12] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *NeurIPS*, 2019.

[13] A. K. Menon, S. Jayasumana *et al.*, "Long-tail learning via logit adjustment," in *ICLR*, 2021.

[14] R. Imam, "Spacenet: An optimally distributed astronomy data," https://www.kaggle.com/datasets/razaimam45/spacenet-an-optimally-distributed-astronomy-data, 2024, accessed 2025-09-09.