Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.36907

Lightweight Vision Models for Galaxy Morphology Classification: A Comparative Study of ResNet, MobileNet, EfficientNet, and MobileViT on Galaxy10 DECaLS

Xiangyu Liao

Faculty of Science and Engineering, Sorbonne University, Paris, France
15002821341@163.com

Abstract. Automated galaxy morphology classification now rests on parameter-heavy convolutional backbones, whose cost becomes a design constraint as survey volumes grow. This paper compares three lightweight architectures (MobileNetV2, EfficientNetB0, and MobileViT-Extra Small (XS)) against a ResNet50 baseline on the Galaxy10 Dark Energy Camera Legacy Survey (DECaLS) benchmark of 17,736 images in ten classes, holding data splits, augmentation, optimizer, loss, and stopping rule fixed so that the backbone is the only variable. The lightweight models give up between 2.07 and 2.55 percentage points of support-weighted accuracy while removing 82.9% to 91.8% of the parameters and 42% to 59% of the per-image inference latency, which makes them 5.7 to 11.8 times more efficient per parameter. Two results run against expectations. The hardest class is not the rarest one. In a redshift-stratified comparison, the only attention-based backbone falls furthest behind the baseline on exactly the smallest and faintest galaxies, so the advantage reported for Vision Transformers does not carry over to a lightweight hybrid. For survey-scale processing, the trade is worth making; for a catalogue of a few thousand galaxies, it is not.

Keywords: galaxy morphology classification, lightweight neural networks, MobileViT, model efficiency, Galaxy10 DECaLS

1. Introduction

A galaxy's shape can signify how it formed. Spiral arms, bars, bulges, and tidal tails all relate to different physical processes, so classifying galaxy morphology has long been a laborious and challenging task in extragalactic astronomy. Galaxy Zoo changed the scale at which galaxies could be classified, producing reliable volunteer classifications for 304,122 galaxies in the Sloan Digital Sky Survey (SDSS), which remain today's reference dataset [1]. Deep learning has now automated classification itself. The Zoobot Bayesian convolutional neural network (CNN) ensemble, trained on Galaxy Zoo votes for images from the DECaLS, provides automated classification with calibrated uncertainties for 314,000 galaxies from deeper DECaLS images. That ensemble is open-source [2].

The Legacy Survey of Space and Time (LSST) at the Rubin Observatory will be cataloguing tens of billions of galaxies and generating 20 TB of raw images every night [3]. At this scale, the cost of running a classifier over each image becomes a design constraint. MobileNetV2 saves computation

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.36907

through depthwise-separable convolutions in inverted residual blocks [4]. EfficientNet scales depth, width, and input resolution jointly [5]. MobileViT embeds lightweight transformer blocks inside a convolutional backbone [6]. Does the accuracy-efficiency trade-off still hold for galaxy images? Galaxies are small, centred, low-contrast objects, and morphology datasets are massively class-imbalanced — in the Galaxy10 DECaLS benchmark, the largest class (round smooth) has 2,645 images against only 334 in the smallest (cigar-shaped smooth), a ratio of nearly eight to one [7]. A lightweight model that does well on average may still fail on those rare classes that happen to be scientifically interesting.

This paper tests the question directly. ResNet50 is trained as a standard baseline alongside MobileNetV2, EfficientNetB0, and MobileViT-XS on Galaxy10 DECaLS, with data splits, augmentation, optimizer, loss, and stopping rule all fixed. Backbone architecture is therefore the only variable. The four models are compared on classification quality (support-weighted accuracy, precision, recall, and F1-score), efficiency (parameter count, inference latency, and accuracy per million parameters), and behaviour, meaning per-class accuracy on the rarest morphology together with Gradient-weighted Class Activation Mapping (Grad-CAM) attention maps. A further question is whether MobileViT-XS's attention layers let it degrade more gracefully on scarce classes.

Three features of this design are not combined in earlier work. Because the four backbones share one protocol, the accuracy each gives up can be read directly against the parameters and latency it saves, rather than against numbers reported under someone else's training budget. The comparison also separates two notions of difficulty that usually travel together: how rare a class is and how small and faint its galaxies are. Each is measured on its own, the second by stratifying the test set by redshift. Grad-CAM maps produced under one identical procedure then show where each backbone actually looks.

2. Related work

2.1. Deep learning for galaxy morphology classification

The state of the art is Walmsley et al., whose ensemble of Bayesian CNNs classifies 314,000 galaxies from deeper DECaLS imaging, released as the open-source Zoobot library [2]. Deep learning has become the new baseline, but computational efficiency is treated throughout as an afterthought.

2.2. Architecture comparison studies and lightweight CNNs

Three standard backbones within the Zoobot framework were compared by Fielding et al., finding DenseNet121 most accurate, but no parameter-normalised metrics were reported [8]. The closest work to this paper is by Wu and others: the EfficientNet-G3 obtained through their architecture search has only 2.17M parameters and an accuracy of 96.63%. Their comparison pool includes MobileNetV2 as the strongest off-the-shelf lightweight model, which is also why it was chosen here [9]. All these studies evaluated their models under their own setups, so a unified comparison across different architecture families is still missing.

2.3. Vision transformers and lightweight hybrid models

Vision Transformers (ViTs) process images as patch sequences under self-attention and compete with CNNs when data are sufficient [10]. Lin et al. first applied one to galaxy morphology. Overall

DOI: 10.54254/2755-2721/2026.36907

accuracy reached 80.55% against a ResNet50 baseline of 85.12%, but the transformer is better on small and faint galaxies [11]. Their evidence is the subset on which the two networks disagree, whose ViT-correct half skews small and faint, rather than accuracy compared inside matched size bins. MobileViT brought such models to mobile scale by embedding transformer blocks in a convolutional backbone [6]. Tan applied it to Galaxy10 DECaLS, reporting over 87% accuracy [12]. Those studies differ in dataset and training budget, so their figures cannot be placed side by side; one identical pipeline for all four models is the gap this paper fills.

3. Methodology

The overall workflow is summarised in Figure 1, which is read from top to bottom and lays out the four stages described in the rest of this section.

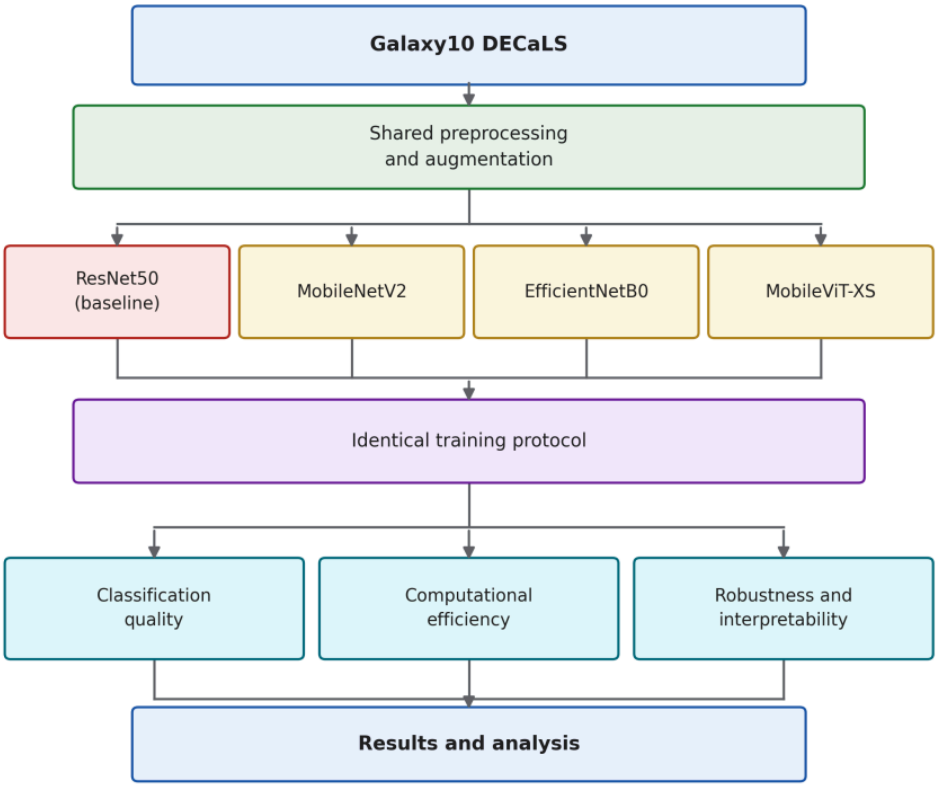

Figure 1. Overall methodological framework of this study (photo credit: original)

The chain starts from the Galaxy10 DECaLS archive and passes through one preprocessing and augmentation stage, drawn once because all four backbones share it. Only in the middle row do the pipelines diverge. ResNet50 in red is the baseline, and the three yellow boxes are the lightweight candidates. Their outputs converge again on a single training protocol, and that is what keeps the comparison controlled. The three boxes below it are the evaluation axes of Section 3.4.

DOI: 10.54254/2755-2721/2026.36907

3.1. Dataset and preprocessing

The dataset is Galaxy10 DECaLS, a set of 17,736 colour images of 256×256×3 pixels covering ten morphology classes, taken from Dark Energy Spectroscopic Instrument (DESI) Legacy Imaging Surveys and labelled on Galaxy Zoo volunteer votes [7]. The class distribution is not uniform: round smooth galaxies provide 2,645 images, and cigar-shaped smooth galaxies just 334. This imbalance is kept, since behaviour under realistic class frequencies is part of what this study measures. The splits are 70% training, 15% validation, and 15% test, stratified so that every class keeps its original proportion in each split. Images are resized to 224×224 to match the ImageNet-pretrained backbones. All three Red, Green, and Blue (RGB) channels are kept rather than being converted to grayscale. The training images are augmented with random flips and rotations of up to 25°, since galaxies have no preferred orientation. The validation and test images are only resized and normalised, with standard ImageNet channel statistics throughout.

3.2. Model architectures

The baseline in this study is ResNet50, which has about 23.5M parameters and represents the standard backbones used for existing morphology pipelines [13]. It is compared against three lightweight models. MobileNetV2, about 2.2M parameters, is built from depthwise-separable convolutions in inverted residual blocks [4]. EfficientNetB0, around 4.0M parameters, is the base model of the compound-scaling family formed from mobile inverted bottleneck convolution blocks [5]. MobileViT-XS, about 2.3M parameters, combines convolution and lightweight transformer blocks for global context [6]. The three lightweight models were chosen for being in the same size range. All the models start with ImageNet-pretrained weights. The classification head is replaced with global average pooling and a fully connected 10-way softmax layer; nothing else in the architecture changes.

3.3. Training protocol

All four models are trained in the same regime: Adam with an initial learning rate of 1×10^{-4} and weight decay of 1×10^{-4}, cross-entropy loss, batch size 64, up to 50 epochs. If validation loss stalls for 3 epochs, the learning rate is halved; if it stalls for 7 consecutive epochs, training stops. The checkpoint with the lowest validation loss is the one tested. Random seeds for splitting and initialisation are fixed in every run, so the experiments are reproduced exactly.

3.4. Evaluation metrics

The evaluation covers three aspects of model behaviour. For classification quality, accuracy, precision, recall, and F1-score are reported for the held-out test set, weighted by class support. Row-normalised confusion matrices show qualitatively where the errors fall. Also reported are the trainable parameter count, single-image inference latency in milliseconds, and accuracy per million parameters.

The third aspect, robustness and interpretability, is the one most specific to this study. Accuracy is broken down by class, in particular, cigar-shaped smooth galaxies, to test whether MobileViT-XS handles sparse data better. Label scarcity is not what Lin et al. tested, though. Their ViT did well on small and faint galaxies [11]. Apparent size is a property of the galaxy itself rather than of its class. A second analysis therefore stratifies the test set by redshift, a proxy for angular size and brightness

DOI: 10.54254/2755-2721/2026.36907

once pixel scale is held fixed, and compares each lightweight model against the ResNet50 baseline inside each bin. Working with the within-bin difference keeps the uneven class composition of the bins from being read as a redshift effect. Grad-CAM maps of the same test images are also created for the four models [14]. Each map is read for one thing. Does the network look at the bar, the arms, the disk edge, or at something in the background that it has no need of?

4. Results and analysis

All experiments were run on a single laptop; the environment is listed in Table 1. Reported latencies and epoch times should be read against that hardware.

Table 1. Hardware and software environment used for all experiments

Component	Specification
Chip / memory	Apple M4, 16 GB unified memory
Compute backend	PyTorch Metal Performance Shaders (MPS)
Python	3.13
PyTorch / torchvision	2.13.0 / 0.28.0
timm	1.0.28
Other libraries	NumPy 2.5.1, scikit-learn 1.9.0, h5py 3.16.0, Matplotlib 3.11.1

4.1. Overall comparison

All results below come from the 2,661-galaxy test set, unseen during training and model selection, and are summarized in Table 2. ResNet50 is the most accurate backbone with 0.8617 weighted accuracy and 0.8615 weighted F1. The three lightweight models are within half a percentage point of each other (EfficientNetB0 at 0.8410, MobileViT-XS at 0.8369, MobileNetV2 at 0.8362). Choosing among them, therefore, costs far less than choosing any of them over the baseline, and even that larger gap is at most 2.55 percentage points. The cost differs sharply. MobileNetV2 uses 2.24M trainable parameters against ResNet50's 23.53M, which is 90.5% fewer, and runs in 3.49 ms per image instead of 8.46 ms; MobileViT-XS is the smallest of the four at 1.94M parameters and 4.76 ms.

Table 2. Test-set performance and computational cost of the four backbones

Model	Accuracy	Precision	Recall	F1	Parameters	Latency (ms)	Accuracy per million params
ResNet50 (baseline)	0.8617	0.8626	0.8617	0.8615	23,528,522	8.46	0.037
MobileNetV2	0.8362	0.8365	0.8362	0.8356	2,236,682	3.49	0.374
EfficientNetB0	0.8410	0.8394	0.8410	0.8375	4,020,358	4.90	0.209
MobileViT-XS	0.8369	0.8340	0.8369	0.8330	1,936,698	4.76	0.432

DOI: 10.54254/2755-2721/2026.36907

4.2. The efficiency–accuracy trade-off

Accuracy per million parameters makes this trade-off explicit: 0.037 for ResNet50 versus 0.432 for MobileViT-XS, 0.374 for MobileNetV2, and 0.209 for EfficientNetB0 — 11.8, 10.2, and 5.7 times the baseline. Whether that is worth paying depends on scale: at a few thousand galaxies, it is not; at the volumes LSST will produce, a 59% cut in per-image latency across tens of billions of objects outweighs two and a half percentage points of accuracy [3]. That latency is measured one image at a time, at batch size 1, on the laptop of Table 1, so the saving will differ on batched server hardware.

4.3. Per-class behaviour and minority-class robustness

Table 3 gives per-class accuracy for five of the ten classes, with n being the number of test images in each, and the pattern is not what a class-imbalance argument would predict. The hardest class for all models is Disturbed (0.623 for ResNet50 and 0.432 for MobileViT-XS) despite its 162 test images, while Cigar Shaped Smooth, with 50, is classified more easily by all four. Scarcity alone, therefore, does not decide difficulty, though Disturbed is itself the second rarest class. Definitional sharpness plausibly matters more: a galaxy is disturbed by degree, and its boundary against a loose spiral or an ongoing merger is genuinely ambiguous. ResNet50 sends 12% of disturbed galaxies to Unbarred Loose Spiral, and MobileViT-XS 25%.

Table 3. Per-class test-set accuracy of the four backbones

Class	n	ResNet50	MobileNetV2	EfficientNetB0	MobileViT-XS
Disturbed	162	0.623	0.525	0.444	0.432
Cigar Shaped Smooth	50	0.880	0.900	0.780	0.700
Unbarred Tight Spiral	274	0.803	0.745	0.810	0.690
Unbarred Loose Spiral	394	0.741	0.756	0.711	0.787
Edge-on with Bulge	281	0.957	0.904	0.957	0.943

For the minority class, the result is not what was expected. MobileNetV2 shows 0.900 on Cigar Shaped Smooth, above ResNet50's 0.880, while MobileViT-XS is the weakest of the four at 0.700 — its errors are at least physically sensible, with 16% going to Edge-on with Bulge, both being elongated, high-inclination systems. When training examples are scarce, the attention-based model is not the one holding up best.

4.4. Small and faint galaxies: a comparison with Lin et al.

Since label scarcity and apparent size are different things, the second was tested directly. Splitting the test galaxies that have a valid redshift at the dominant pixel scale into quartiles of 637 gives bins whose median redshift is 0.035 to 0.128, from nearby and well-resolved to distant, small, and faint. The absolute accuracy of each bin is not comparable, since the class mix of the galaxies changes dramatically with redshift: Edge-on without Bulge falls from 22% of the lowest quartile to 0.8% of the highest. The difference from ResNet50 within each bin is therefore used, where all four models see the same galaxies, together with a composition-standardised accuracy that weights each bin to the overall class distribution. Only one of the two is drawn: Figure 2 plots the raw, unstandardised within-quartile difference, and the standardised figures are quoted below.

DOI: 10.54254/2755-2721/2026.36907

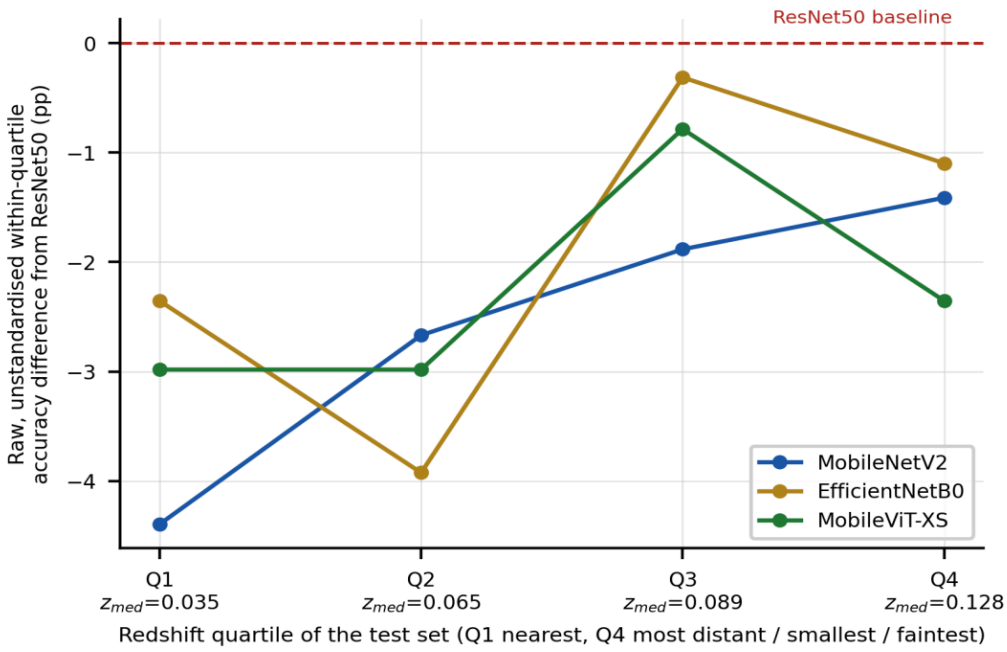

Figure 2. Raw accuracy difference from the ResNet50 baseline across redshift quartiles of the test set (photo credit: original)

The effect Lin et al. report would show up in Figure 2 as a MobileViT-XS curve climbing towards the right. It does not. Its raw difference from ResNet50 is −2.98, −2.98, −0.78, and −2.35 percentage points from Q1 to Q4, with no trend. The standardised results below are quoted as shortfalls, that is, the percentage points by which a model classified fewer of a quartile's galaxies correctly than ResNet50 did, so a smaller shortfall is closer to the baseline and a negative one would be ahead of it. MobileViT-XS gives 2.7, 2.7, 1.9, and 5.3, with its Q4 value being the largest shortfall recorded for any model in any quartile. MobileNetV2 gives 4.7, 2.6, 2.2, and 1.3, shrinking monotonically as redshift rises. MobileViT-XS is the only backbone here with attention, so it alone bears on Lin et al.'s finding; the two convolutional models are controls. Read as a design rule, that attention holds up better where galaxies are smallest and faintest, the finding fails. In Q4, the attention-based backbone is the worst of the four at 5.3; both convolutional models stay below 1.9, and no shortfall is negative anywhere. None of this contradicts Lin et al. themselves. Their transformer, dataset, training budget, and measure of galaxy size all differ from those used here. The generalisation drawn from their result fails; the result itself stands. Whether MobileViT-XS is undertrained, being the slowest to converge at 31 epochs on a dataset nearly nine times smaller, or whether the advantage belongs to that architecture alone, cannot be settled here.

4.5. Attention maps

Figure 3 places Grad-CAM maps for the four models on the same four test galaxies, one per row and one backbone per column, with the label under each overlay recording whether that backbone was correct [14]. The three convolutional models are similar: heat concentrates on the galaxy and follows the structure the label depends on, leaving the surrounding sky cool.

DOI: 10.54254/2755-2721/2026.36907

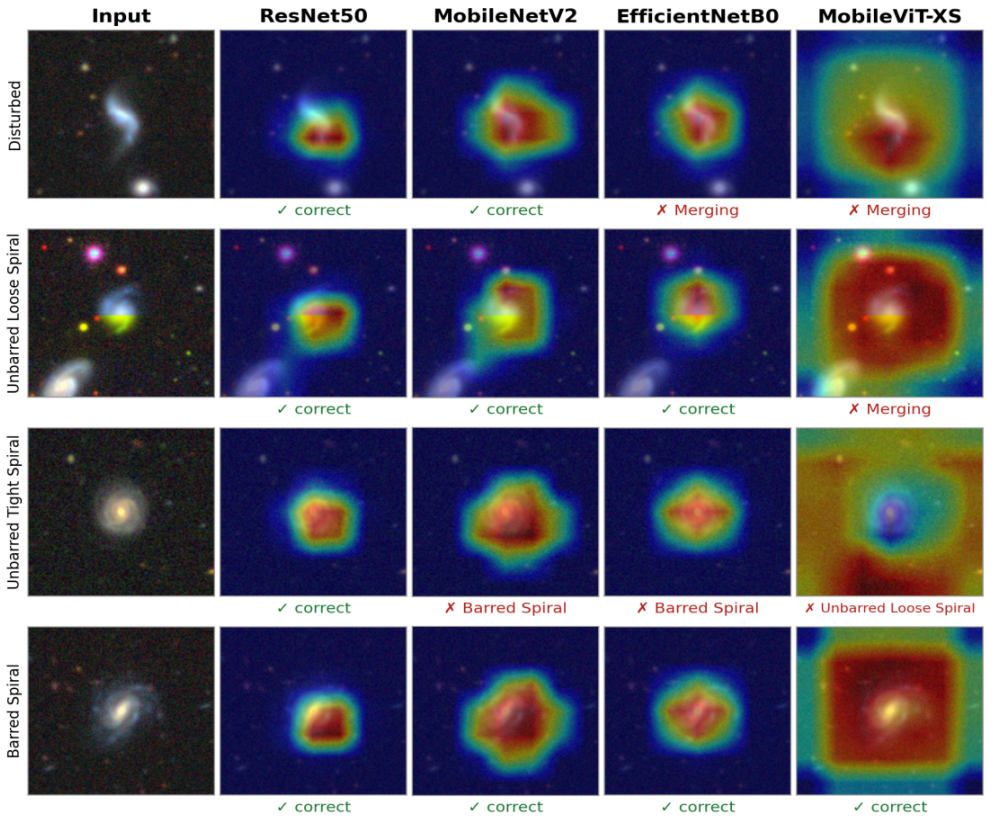

Figure 3. Grad-CAM attention of the four backbones on four representative test galaxies (photo credit: original)

MobileViT-XS is different. Its maps are much more diffuse, and in two of the four examples, the strongest response is in the background sky, and the galaxy stays cool. That is not an artefact of where the hook is placed. The layer used is the backbone's final 1×1 convolution, the direct counterpart of the layer used for ResNet50 and of the same resolution. It still does not follow that the model is reading the sky. Grad-CAM assumes a feature-map position corresponds to an image position, and self-attention breaks that assumption by mixing information across all positions. The contrast is reported as an observation, not a causal claim; settling it would require an attribution method built for attention-based models.

5. Conclusion

ResNet50 was compared against MobileNetV2, EfficientNetB0, and MobileViT-XS on Galaxy10 DECaLS under a single training protocol in which the backbone was the only variable. The lightweight models cost between 2.07 and 2.55 percentage points of weighted accuracy while removing between 82.9% and 91.8% of the parameters and between 42% and 59% of the inference latency, making them 5.7 to 11.8 times more efficient per parameter. For survey-scale processing, that is a trade worth making; for a small sample, it is not.

Two findings came as a surprise. Class rarity alone does not decide difficulty. Disturbed, with 162 test images, was classified worse by every model than Cigar Shaped Smooth, with 50. Nor did the attention-based backbone buy robustness where it was expected. Read as a design rule, the advantage Lin et al. report on small and faint galaxies does not carry over. MobileViT-XS, the only

DOI: 10.54254/2755-2721/2026.36907

attention-based backbone tested, fell furthest behind the baseline on exactly the faintest galaxies. Of the lightweight models, the purely convolutional MobileNetV2 was the more robust choice on both axes examined here, despite MobileViT-XS being the smaller network.

The limits are clear. One dataset, one training budget, and one seed per model leave the two- to three-point differences here without an estimate of run-to-run variance. MobileViT-XS may be penalised by a schedule that suited the convolutional models better; redshift is only a proxy for angular size and brightness; and the Grad-CAM comparison is qualitative and, for the hybrid model, of uncertain validity. Repeated runs with different seeds, a longer schedule for the transformer-based model, and an attribution method suited to attention are the natural next steps.

References

[1] Willett, K. W., Lintott, C. J., Bamford, S. P., et al. (2013). Galaxy Zoo 2: Detailed morphological classifications for 304, 122 galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society, 435*, 2835–2860. https://doi.org/10.1093/mnras/stt1458

[2] Walmsley, M., Lintott, C., Géron, T., et al. (2022). Galaxy Zoo DECaLS: Detailed visual morphology measurements from volunteers and deep learning for 314, 000 galaxies. *Monthly Notices of the Royal Astronomical Society, 509*, 3966–3988. https://doi.org/10.1093/mnras/stab2093

[3] Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. (2019). LSST: From science drivers to reference design and anticipated data products. *The Astrophysical Journal, 873*, 111. https://doi.org/10.3847/1538-4357/ab042c

[4] Sandler, M., Howard, A., Zhu, M., et al. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4510–4520). Salt Lake City, USA. https://doi.org/10.1109/CVPR.2018.00474

[5] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 6105–6114). Long Beach, USA.

[6] Mehta, S., & Rastegari, M. (2022). MobileViT: A lightweight, general-purpose, and mobile-friendly vision transformer. In *Proceedings of the International Conference on Learning Representations*.

[7] Leung, H. W., & Bovy, J. (2024). Galaxy10 DECaLS [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.10845026

[8] Fielding, E., Nyirenda, C. N., & Vaccari, M. (2021). A comparison of deep learning architectures for optical galaxy morphology classification. In *Proceedings of the 2021 International Conference on Electrical, Computer and Energy Technologies* (pp. 1–5). Cape Town, South Africa. https://doi.org/10.1109/ICECET52533.2021.9698414

[9] Wu, D., Zhang, J., Li, X., et al. (2022). A lightweight deep learning framework for galaxy morphology classification. *Research in Astronomy and Astrophysics, 22*, 115011. https://doi.org/10.1088/1674-4527/ac92f7

[10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*.

[11] Lin, J. Y.-Y., Liao, S.-M., Huang, H.-J., et al. (2021). Galaxy morphological classification with an efficient vision transformer. In *Proceedings of the Fourth Workshop on Machine Learning and the Physical Sciences, NeurIPS*.

[12] Tan, X. (2024). Accurate and efficient galaxy classification based on a mobile vision transformer. *Applied Computing and Engineering, 33*, 118–125. https://doi.org/10.54254/2755-2721/33/20230245

[13] He, K., Zhang, X., Ren, S., et al. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). Las Vegas, USA. https://doi.org/10.1109/CVPR.2016.90

[14] Selvaraju, R. R., Cogswell, M., Das, A., et al. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626). Venice, Italy. https://doi.org/10.1109/ICCV.2017.74