Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Empirical Investigation of Vision Architectures for High- Cardinality Fine-Grained Ancient Chinese Ideograph Classification on OBWX

Yanqiang Ge

gyq@aynu.edu.cn

Anyang Normal University

Zixu Yang

Ministry of Education of China, Anyang Normal University

Ziqi Ge

Jilin University

Zhan Zhang

Ministry of Education of China, Anyang Normal University

Haoyu Wang

Ministry of Education of China, Anyang Normal University

Article

Keywords: ancient Chinese character recognition, fine-grained classification, benchmark evaluation, model comparison, long-tail distribution

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11430158/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Empirical Investigation of Vision Architectures for High-Cardinality Fine-Grained Ancient Chinese Ideograph Classification on OBWX

Yanqiang Ge[1,2], Zixu Yang[1,2], Ziqi Ge[3], Zhan Zhang[1,2], Haoyu Wang[1,2]

[1] School of Computer and Information Engineering, Anyang Normal University, Anyang, China

[2] Key Laboratory of Oracle Bone Inscriptions Information Processing, Ministry of Education of China, Anyang Normal University, Anyang, China

[3] School of Computer Science and Technology, Jilin University, Changchun, Jilin, China

Corresponding author: Yanqiang Ge (gyq@aynu.edu.cn)

Abstract：
Ancient Chinese ideograph recognition is a high-cardinality fine-grained task featuring thousands of visually similar categories and a severely long-tailed data distribution. Existing studies adopt divergent dataset construction, class cardinality, and evaluation protocols, rendering their reported accuracies not directly comparable. We build a controlled empirical benchmark by evaluating 14 backbones from five architectural families with all evaluation variables held constant on the public OBWX dataset (74,047 glyph images, 6,650 categories). To disentangle the confounding effect of class frequency on single overall accuracy metrics, results are reported along four dimensions: computational cost, accuracy stratified by sample-frequency bands, 95% Wilson confidence intervals on the 12,583-sample test set, and performance degradation under controlled noise injection. EfficientNetV2-S achieves the highest accuracy (74.51%), outperforming hybrid MaxViT-T (72.31%) and hierarchical Swin Transformer (71.57%). Under the primary seed, classical CNNs lag substantially at 36.3% – 36.7%, with a 32.09 percentage point (pp) performance span across architecture families. Neither parameter count nor FLOPs correlates with recognition accuracy; the largest model, ViT-B/16, underperforms by 9.07 pp despite having fourfold parameters. Accuracy decreases monotonically from head (90.32%) to tail (60.44%) categories, with tail classes accounting for 64.0% of all errors. Noise injection reverses the clean-data ranking: MaxViT-T retains 83.0% of its clean accuracy, compared with 77.4% for EfficientNetV2-S and 68.5% for Swin. Three-seed replication validates the top ranking while identifying three unstable backbones with 13.7 – 24.8 pp performance variations.

Keywords： ancient Chinese character recognition; fine-grained classification; benchmark evaluation; model comparison; long-tail distribution

1 Introduction

1.1 Background

Ancient Chinese characters are the most direct and most stable written remains of early Chinese civilisation, and their graphic evolution follows a broad trajectory from concrete depiction toward abstraction and from irregular complexity toward standardisation [1], [2]. The span from oracle bone script to small seal script marks the transition in which the writing system took shape. Scholars have identified roughly 4,500

character heads since the discovery of oracle bone inscriptions, yet only about 1,500 have been deciphered [1], [2]. Collating and retrieving the undeciphered characters and graphic variants remains specialised, time-consuming, and dependent on individual expertise. Automatic classification at the character-head level has become a foundational capability for digital collation and computer-assisted decipherment: a single character head appears in markedly different forms across periods, media, and regions, so a model must discriminate precisely within several thousand categories. Computational palaeography has recently emerged at the intersection of artificial intelligence and palaeography [3], moving from isolated automation toward integrated digital ecosystems—which makes standardised, reproducible benchmarks an urgent requirement.

1.2 Overview of Recent Progress

On the data side, high-quality open datasets have advanced the field. HUST-OBC [4] provides 140,053 oracle bone images, of which 77,064 show deciphered characters (1,588 heads) and 62,989 undeciphered ones, with a ResNet baseline at 94.6% accuracy and 0.914 macro-averaged F1. EVOBC [5] covers 229,170 images and 13,714 heads across six historical stages. OBWX [6], released in 2026 and drawn from four authoritative palaeographic compilations, covers oracle bone, bronze, Warring States, and small seal scripts with 74,047 images and 6,650 categories. In recognition methodology, generative and reconstructive paradigms have been representative since 2024: OBSD maps undeciphered characters onto modern Chinese character images with a conditional diffusion model [7]; P3 reconstructs modern equivalents through radical decomposition and releases the ACCP dataset over seven stages [8]; BIRD addresses restoration and dating of bronze inscriptions [9]; and a decoupled method targets long-tailed distributions [10]. Large multimodal models have since entered the field, with OBI-Bench [11] and V-Oracle [12] evaluating reasoning beyond classification. Diffusion-based generation and structurally constrained glyph synthesis continue to extend this line [13], [14]. Two recent surveys systematise the field's data resources, method lineages and open challenges from the perspectives of character recognition [15] and complete information-processing pipelines [16].

1.3 Open Problems and Motivation

Despite these advances, the evaluative basis for fine-grained character classification remains weak. First, existing public datasets either concentrate on a single period or rely on web-crawled images of uneven quality; none offers a large-category benchmark from authoritative publications that covers consecutive periods and is systematically verified. Second, studies differ in splitting strategy, augmentation policy, and training length, so reported figures are not comparable and backbone selection lacks a common reference. Third, categories follow a pronounced long tail, and head and tail classes place different demands on a model, yet no systematic evaluation of mainstream architectures under such a distribution exists. Fourth, the convolutional-versus-transformer competition has been discussed in this domain, but quantitative evidence under cross-period, large-category conditions remains limited. We address these gaps with a unified benchmark on OBWX.

1.4 Contributions

Using OBWX [6], we build a unified evaluation benchmark for cross-period fine-grained ancient character classification (dataset details in Section 3.1). The main contributions are as follows.

(1) A unified, fairness-first evaluation protocol and metrics system. Every model is trained and tested under strictly identical conditions: stratified partitioning at a nominal per-category 70%/15%/15% ratio, fidelity-

constrained augmentation, long-tail mitigation, and sample-weighted metrics. Results are therefore directly comparable and reproducible.

(2) A systematic multi-family comparison. The fourteen architectures belong to five families: classical CNNs (ResNeXt-50, WideResNet-50), lightweight CNNs (MobileNetV3-Large, RepViT-M1.1), CNN–Transformer hybrids (MaxViT-T, EfficientFormerV2-S, MobileViT-V2-S), modern ConvNets (EfficientNetV2-S, EfficientNet-B3, InceptionNeXt-Small, ConvNeXt-V2-Tiny), and pure Transformers (ViT-B/16, Swin Transformer, PVTv2-B2). All are evaluated under the same protocol, yielding a complete performance ranking over 6,650 categories together with a family-level analysis of accuracy, cost, and error structure.

(3) A scientific error and difficulty analysis. Beyond the accuracy ranking, we decompose errors by category-frequency band and confusion pattern and quantify the long-tail effect (tail categories contribute 64.0% of errors at $1.55\times$ error density). The core bottleneck is component-level discrimination of visually similar heads, not diachronic variation: among the 3,210 errors of the best model under the per-sample prediction set used for the band analysis (Section 4.4), none is a same-head error, although the base for that statistic is narrow because only 111 test samples, distributed over 34 character heads, admit same-head confusion at all; across the twelve analysed models no single model produces more than three such errors, never exceeding 0.06% of its errors.

(4) Reusable evaluation insights. Every figure carries a 95% confidence interval over the 12,583-sample test set, and Section 4.7 repeats the full benchmark under three seeds (42, 123, 2024). The replication holds at the head of the ranking: the leading model and the three-backbone leading set are identical under every seed, whereas three unstable backbones show seed spreads of 13.7–24.8 pp. Model selection by validation accuracy also transfers to the test set, reproducing 12 of the 14 ranking positions. Family means span 32.09 pp, and the two extreme families separate from the rest by 11.7–14.5 pp; Section 4.2 shows, however, that among the three design-contemporary families the ordering is not supported ($\eta^2 = 0.11$, permutation p = 0.69). A noise-injection sweep over the three leading backbones (Section 4.6) shows that the clean-set ordering does not survive degradation. These results give practitioners transferable guidance on backbone selection.

2 Related Work

2.1 Ancient Script Datasets

Data underpins any intelligent processing of ancient scripts. Early studies relied on small, individually assembled collections of rubbings with inconsistent scale and annotation conventions. Table 1 summarises the representative public datasets. Oracle-20K [17] and OBC306 [18] established early benchmarks; with 309,551 rubbing images, OBC306 was the largest resource of its time, though its category distribution is highly imbalanced. Oracle-AYNU [19] and HWOBC [20] extended handwritten samples to tens of thousands, and Oracle-MNIST [21] provides a lightweight benchmark for robustness to noise and deformation. EVOBC [5] and ACCP [8] broaden the scope to cross-stage graphic evolution, although part of EVOBC is web-crawled and variable in quality. HUST-OBC [4] is among the largest open oracle bone datasets with over 140,000 images, while a graph-structured dataset pioneered evolution-chain modelling with 756 chains [22].

Table 1 Comparison of major public ancient script datasets

Dataset	Year	Coverage	Images	Categories	Main characteristics
Oracle-20K [17]	2016	Oracle bone	20,039	261	Handwritten oracle bone; hierarchical representation learning benchmark
OBC306 [18]	2019	Oracle bone	309,551	306	Rubbing images; highly imbalanced category distribution
Oracle-AYNU [19]	2019	Oracle bone	39,062	2,583	Handwritten oracle bone; pronounced long tail
HWOBC [20]	2020	Oracle bone	83,245	3,881	Handwritten oracle bone; multiple writing styles
Oracle-MNIST [21]	2024	Oracle bone	30,222	10	MNIST-compatible format; noise and deformation robustness benchmark
EVOBC [5]	2024	Oracle bone – clerical (6 stages)	229,170	13,714	Targets graphic evolution; some images web-crawled
HUST-OBC [4]	2024	Oracle bone	140,053	1,588 + 9,411 (categories)	Deciphered/undeciphered system; semi-automatic construction pipeline
OBWX [6]	2026	Oracle bone – small seal (4 periods)	74,047	6,650	Authoritative compilation source; diachronic labels; expert verified

Existing resources fall short in one of two ways: most are confined to a single period [17]–[21], and the rest lean on web images whose quality varies [5]. None provides a fine-grained benchmark from authoritative publications that covers consecutive periods and is systematically verified. To address this gap, our group constructed OBWX [6] from four authoritative palaeographic compilations [23]–[26], spanning oracle bone, bronze, Warring States, and small seal scripts. Its construction pipeline, source details, and expert-verification procedure are documented in the dataset paper [6] and summarised in Section 3.1.

2.2 Ancient Character Recognition Methods

Research on ancient character recognition has moved from hand-crafted features to deep feature learning. Methods for general handwritten Chinese character recognition supply the methodological basis [27], while progress in document intelligence has supported the digitisation of ancient scripts [28]. At the architectural level, residual networks [29], densely connected networks [30], and vision transformers [31] have become mainstream backbones. These learn discriminative deep representations rather than prescribed stroke or contour features, accommodating the variable glyph shapes, diverse media, and co-occurrence of rubbing damage and stroke loss typical of ancient scripts.

Studies specific to ancient scripts include the following. Zhou et al. [32] curated a bronze inscription dataset of 198,598 annotated characters across 6,658 categories and equipped a two-stage detection–recognition pipeline with mixture-of-experts adapters on a pretrained CLIP encoder. Wang et al. [33] proposed a multi-modal framework with cross-modal homogenisation that handles hybrid inputs such as rubbings and handwritten copies. Component-level modelling has also been explored: HCT-Net [34] aggregates local and global branches through gated fusion with a spectral filter, and cross-modal metric learning aligns glyph images with radical text [35]. Wu et al. [36] addressed open-set conditions with

Dynamic GridMask and intra-class fusion, and Ge et al. [37] combined a deep autoencoder with a fuzzy classifier for interpretable recognition of degraded characters. Since 2024 the field has extended toward integration and long-tail robustness: OBSD [7] generates decipherment cues with conditional diffusion, P3 [8] performs cross-period decipherment through radical reconstruction, BIRD [9] addresses bronze inscription restoration and dating, and a decoupled strategy targets long tails [10]. Prism-OBI decouples perception from semantic interpretation [38], component-level segmentation combines sparse strong with weak annotations and expert priors [39], Oracle-P15K supplies a structure-aligned dataset with the OBIDiff generator [40], and hybrid convolutional–transformer designs alleviate imbalance [41]. The repertoire has expanded from discriminative classification to generation, reconstruction, and cross-modal reasoning, yet a systematic quantitative evaluation of mainstream backbones for fine-grained classification across periods and under long tails is still missing—and that is what this benchmark supplies.

2.3 Deep Learning Backbones

Backbone choice determines both the accuracy ceiling and the deployment cost. We organise the fourteen architectures into five families marking successive design vintages. (i) Classical CNNs: AlexNet [42] and VGG [43] established deep convolution; ResNet [29] introduced residual connections and DenseNet [30] dense connectivity; ResNeXt [44] and WideResNet-50 [45] extend the residual design along cardinality and width and are the two classical representatives here. (ii) Lightweight CNNs: MobileNetV2 [46] uses inverted residuals, MobileNetV3 [47] adds architecture search, ShuffleNetV2 [48] uses channel shuffling, and RepViT [49] revisits mobile CNNs from the ViT perspective. (iii) Modern ConvNets: EfficientNet [50] introduces compound scaling, EfficientNetV2 [51] adds Fused-MBConv and progressive training, RegNet [52] uses design-space search, ConvNeXt [53] reformulates the pure ConvNet under modern principles, ConvNeXt V2 [54] adds a fully convolutional masked autoencoder, and InceptionNeXt [55] decomposes large-kernel depthwise convolution. (iv) CNN–Transformer hybrids: CoAtNet [56] unifies depthwise convolution and relative attention, MaxViT [57] uses multi-axis attention, MobileViT-V2 [58] reduces attention to linear complexity, and EfficientFormerV2 [59] attains MobileNet-level latency through a searched supernet. (v) Pure transformers: ViT [31] introduced isotropic patch-token self-attention, Swin Transformer [60] uses a shifted-window hierarchy, and PVT v2 [61] adds linear-complexity attention and overlapping patch embedding.

3 Benchmark Configuration

3.1 The OBWX Dataset

This paper adopts the OBWX dataset of cross-period ancient Chinese ideographs [6]. It contains 74,047 glyph images organised into 6,650 ideographic character categories, covering four historical periods: oracle bone, bronze, Warring States, and small seal scripts. Its glyph images are drawn from four authoritative palaeographic compilations and underwent glyph cropping, quality filtering, and verification by three specialist scholars; its category distribution has a pronounced long tail (Gini coefficient 0.6693, computed from the released per-category image counts). Table 2 reports the per-period composition, while the complete construction pipeline, source details, statistical properties, and release licence are given in Ref. [6].

Table 2 Composition of the OBWX dataset across the four periods

Period	Historical range	Categories	Images	Typical media
Oracle bone	Shang dynasty (c. 14th–11th c. BCE)	1,848	6,204	Turtle plastrons, animal bones
Bronze	Shang and Zhou dynasties	3,339	22,388	Bronze vessels
Warring States	475–221 BCE	5,003	41,040	Bamboo/silk manuscripts, currency, seals
Small seal	Qin dynasty (after 221 BCE)	4,031	4,415	Stone inscriptions, imperial edict plates
Total	—	6,650 (unique character heads after de-duplication across periods)	74,047	—

3.2 Data Partitioning and Augmentation

The core of the benchmark protocol is to train and test all models under strictly identical conditions; the pipeline is shown in Fig. 1. Partitioning is stratified: images are split into training, validation, and test sets at a nominal per-category ratio of 70%, 15%, and 15%, preserving the natural distribution of periods and categories. Because categories with at most three images go entirely to training and the rest are split by integer allocation, the realised split deviates from the nominal ratio: 118,359 training images (after the offline tail augmentation below), 9,540 validation, and 12,583 test. Such categories are assigned wholly to training so that the model has seen every category at least once; validation and test contain no samples of them. The training set therefore covers all 6,650 categories, whereas validation and test cover 3,447 (51.8%); the remaining 3,203 do not participate in evaluation. This guarantees that every category is seen during training, but the reported metrics exclude a generalisation assessment of the rarest categories, which must be examined separately with few-shot methods.

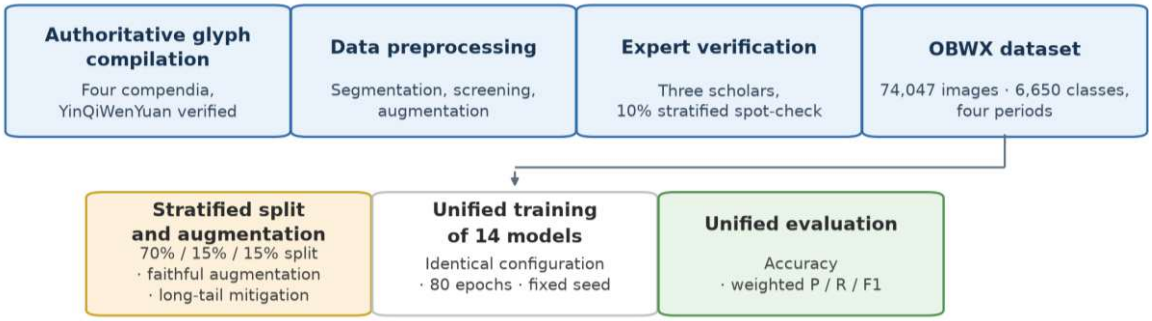


Fig. 1 Overall pipeline of the OBWX fine-grained character classification benchmark

Augmentation follows a fidelity constraint: fine-grained character-head discrimination demands greater glyph fidelity than coarse-grained tasks, since aggressive geometric transformations destroy glyph structure and horizontal flipping produces historically untenable orientations. Training augmentation is therefore restricted to two operations, each applied independently with probability 0.3: random rotation within ±5° (constant fill, value 255) and brightness/contrast jitter within ±10%; flipping is forbidden. To mitigate the long tail, categories whose training portion holds fewer than 15 images are augmented offline to 15 using minimal transformations (±1° rotation, ±2% scaling, ±0.5° shearing, ±1% translation), applied after longest-side resizing to 256 px, padding to 256 × 256 and random cropping to 224 × 224; the augmented images are generated once before training and held fixed. Validation and test receive no augmentation. The unified

training configuration is given in Table 3.

Table 3 Unified training configuration of the benchmark experiments

Configuration item	Setting
Input size	224 × 224
Normalisation	ImageNet mean / standard deviation
Batch size	64
Training epochs	80
Random seed	42 (primary); 123 and 2024 (three-seed replication, Section 4.7)
Data augmentation	Random rotation ±5° and brightness/contrast jitter ±10%, each applied independently with probability 0.3; horizontal flipping disabled
Long-tail mitigation	Categories with fewer than 15 images augmented offline to 15 (±1° rotation, ±2% scaling, ±0.5° shear, ±1% translation; longest-side resize to 256 px, pad to 256 × 256, then random crop to 224 × 224)
Model selection	Checkpoint with highest validation accuracy used for testing
Partitioning	Stratified sampling, nominal 70%/15%/15% per category; categories with at most three images assigned entirely to training; realised split 118,359 / 9,540 / 12,583 (train count after augmentation)

3.3 Evaluation Metrics

The benchmark adopts accuracy together with precision, recall, and F1 weighted by per-category support. Let N denote the total number of test samples and N_correct the number correctly classified; accuracy is defined in Eq. (1). For the i-th category (of C = 6,650 categories in total; categories without any test sample contribute zero weight and are excluded by construction), let n_i be its number of test samples and P_i and R_i its precision and recall, as given in Eqs. (2) and (3). Weighted metrics average per-category values with the proportion of test samples n_i/N as weight, as shown in Eqs. (4) and (5).

$$Accuracy = \frac{N_{correct}}{N} \quad (1)$$

$$P_i = \frac{TP_i}{TP_i + FP_i} \quad (2)$$

$$R_i = \frac{TP_i}{TP_i + FN_i} \quad (3)$$

$$P_w = \sum_{i_1}^{C} \frac{n_i}{N} P_i \; R_w = \sum_{i_1}^{C} \frac{n_i}{N} R_i \quad (4)$$

$$F1_w = \frac{2P_w R_w}{P_w + R_w} \quad (5)$$

Weighted rather than macro averaging is chosen because, with 6,650 categories under a strong long tail, macro averaging would give tail categories holding one or two samples the same weight as head categories occurring hundreds of times, letting noise from extremely rare categories dominate. The weighted scheme keeps each category's contribution proportional to its sample count, while the rarest categories (at most three original samples, all in training) are excluded by construction. The two conventions answer different questions: weighted metrics measure overall performance by sample count, whereas macro averaging measures whether every character head can be recognised stably; under a strong long tail, their divergence is itself a measure of long-tail impact. We therefore report overall performance with weighted metrics in Section 4.1 and supplement them with macro-averaged recall in Section 4.4.

3.4 Experimental Environment and Implementation

All fourteen models share a single set of data loading, training, and evaluation code; only the network definition and input–output dimensions differ, minimising bias from implementation differences. Every model is trained for the full 80 epochs to ensure adequate convergence on the 6,650-class task, and the checkpoint with the highest validation accuracy is retained for final testing. Reproducibility is a basic requirement of benchmark work; the complete software and hardware configuration is listed in Table 4.

Table 4 Experimental environment configuration

Configuration item	Specification
Operating system	Ubuntu 22.04 LTS
CPU	Intel(R) Xeon(R) Silver 4314 @ 2.40 GHz (64 threads, GenuineIntel)
Memory	256 GB
GPU	6 × NVIDIA GeForce RTX 4090 (24 GB per card)
Programming language	Python 3.x
Deep learning framework	PyTorch 2.0.0 / torchvision 0.15.0; other dependencies: albumentations≥1.3.0, opencv-python≥4.8.0, pillow≥10.0.0, numpy≥1.24.0, pandas≥2.0.0, scikit-learn≥1.3.0, matplotlib≥3.7.0, tqdm≥4.65.0, pyyaml≥6.0
CUDA / cuDNN	12.4; NVIDIA driver version 550.100
Model implementation	torchvision.models and timm (ImageNet pre-trained weights; for every architecture the original classifier is replaced by a unified 6,650-way MLP head)

4 Benchmark Results and Analysis

4.1 Overall Performance Comparison

Table 5 reports the performance of the fourteen models on the OBWX test set, Fig. 2 compares their test accuracies directly, and Fig. 3 shows the training dynamics of the best model, EfficientNetV2-S, and the runner-up, MaxViT-T, together with an item-by-item comparison of five metrics. Reported recall is the sample-weighted recall, which by construction equals overall accuracy; it is retained alongside precision and F1 for completeness.

Table 5 Performance comparison of fourteen models on the OBWX dataset

Model	Family	Test accuracy (%)	Best validation accuracy (%)	Precision	Recall	F1
EfficientNetV2-S	Modern ConvNet	74.51	75.90	0.7347	0.7451	0.7234
MaxViT-T	CNN-Transformer Hybrid	72.31	74.03	0.7100	0.7231	0.6981
Swin Transformer	Pure Transformer	71.57	73.75	0.7036	0.7157	0.6931
PVTv2-B2	Pure Transformer	68.75	70.41	0.6754	0.6875	0.6638
EfficientFormerV2-S	CNN-Transformer Hybrid	66.14	68.63	0.6415	0.6614	0.6314
EfficientNet-B3	Modern ConvNet	65.59	67.25	0.6339	0.6559	0.6242
ViT-B/16	Pure Transformer	65.44	67.37	0.6442	0.6544	0.6291
InceptionNeXt-Small	Modern ConvNet	62.51	65.20	0.6037	0.6251	0.5941
ConvNeXt-V2-Tiny	Modern ConvNet	59.86	61.91	0.5787	0.5986	0.5687
MobileNetV3-Large	Lightweight CNN	51.21	54.14	0.4817	0.5121	0.4792
MobileViT-V2-S	CNN-Transformer	49.54	52.63	0.4678	0.4954	0.4606

Model	Family	Test accuracy (%)	Best validation accuracy (%)	Precision	Recall	F1
RepViT-M1.1	Hybrid Lightweight CNN	45.18	48.71	0.4222	0.4518	0.4180
WideResNet-50	Classical CNN	36.68	40.74	0.3429	0.3668	0.3384
ResNeXt-50	Classical CNN	36.31	40.07	0.3369	0.3631	0.3339

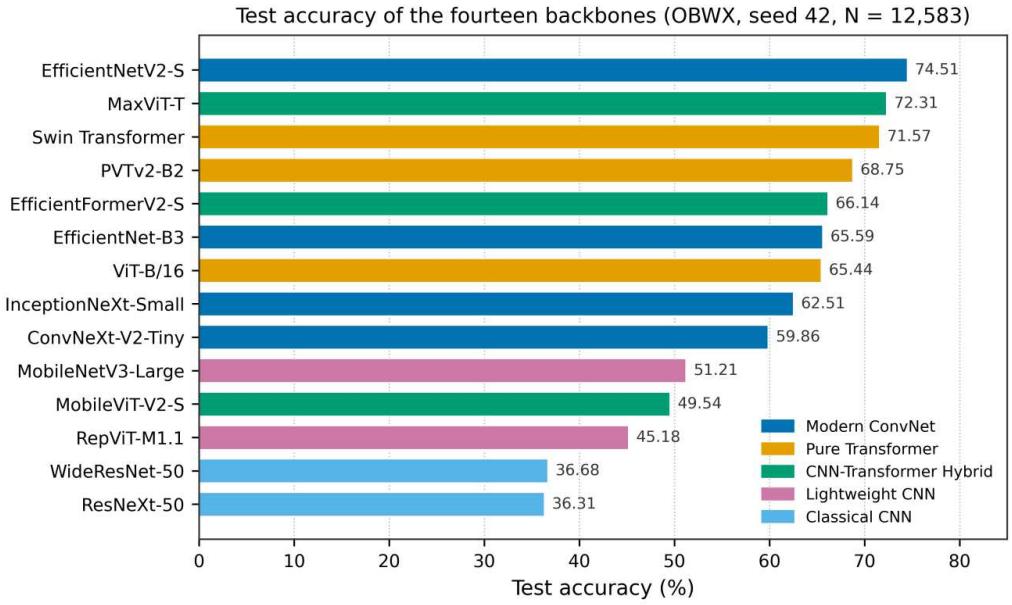


Fig. 2 Test accuracy comparison of the fourteen models (coloured by architectural family)

The results fall into a clear hierarchy. EfficientNetV2-S attains the global optimum (74.51%) and leads on all five metrics (best validation accuracy 75.90%, precision 0.7347, recall 0.7451, F1 0.7234). One reading consistent with these results, offered as a hypothesis rather than an established cause because no ablation isolates individual design choices, is that its Fused-MBConv-centred early design balances the efficiency of local stroke feature extraction against receptive-field growth while its non-uniform stage-wise scaling preserves feature discriminability in the 6,650-category space. The hybrid MaxViT-T follows 2.20 pp behind (72.31%), ranking second in precision (0.7100) and F1 (0.6981); multi-axis attention, which models local stroke detail and global glyph layout jointly, is a plausible contributor, likewise untested by ablation. The hierarchical Swin Transformer (71.57%) and the pyramid PVTv2-B2 (68.75%) complete the leading group, with confidence intervals given in Section 4.5. Below them, EfficientFormerV2-S (66.14%), EfficientNet-B3 (65.59%), and the isotropic ViT-B/16 (65.44%) form the upper-middle band; InceptionNeXt-Small (62.51%) and ConvNeXt-V2-Tiny (59.86%) occupy the middle; the lightweight CNN MobileNetV3-Large (51.21%), the lightweight hybrid MobileViT-V2-S (49.54%), and RepViT-M1.1 (45.18%) form the lower-middle band; and the two classical CNNs, WideResNet-50 (36.68%) and ResNeXt-50 (36.31%), sit at the bottom. The best-to-worst span is 38.20 percentage points; its architectural-family structure is analysed in Section 4.2.

Fig. 3 further characterises the training process and metric comparison of the top two models. In convergence speed EfficientNetV2-S holds a clear advantage: its validation accuracy passes 55% at epoch 5 and reaches 66.23% at epoch 15, by which point training loss has fallen to 1.60. MaxViT-T reaches the corresponding levels noticeably later, with a slower early phase. Across the full run, both models improve steadily from epoch 20 onward: EfficientNetV2-S peaks at 75.97% validation accuracy at epoch 70, the

maximum of the recorded epoch-wise validation curve and 0.07 pp above the 75.90% attached to the retained checkpoint reported in Table 5. MaxViT-T peaks later, at 74.03% on epoch 78. From epoch 20 to 80 the EfficientNetV2-S curve leads by 1.65 percentage points on average. By the end of training both models approach their training-accuracy ceiling (final training losses of 1.355 and 1.348), indicating that the performance difference does not stem from insufficient training but from an intrinsic difference in generalisation. The five-metric comparison in Fig. 3 further shows that EfficientNetV2-S leads MaxViT-T on test accuracy, best validation accuracy, precision, recall, and F1 alike (by 1.87 to 2.53 percentage points), and that its validation and test performance agree closely (75.90% versus 74.51%, Table 5), confirming its robustness as the best model in this benchmark.

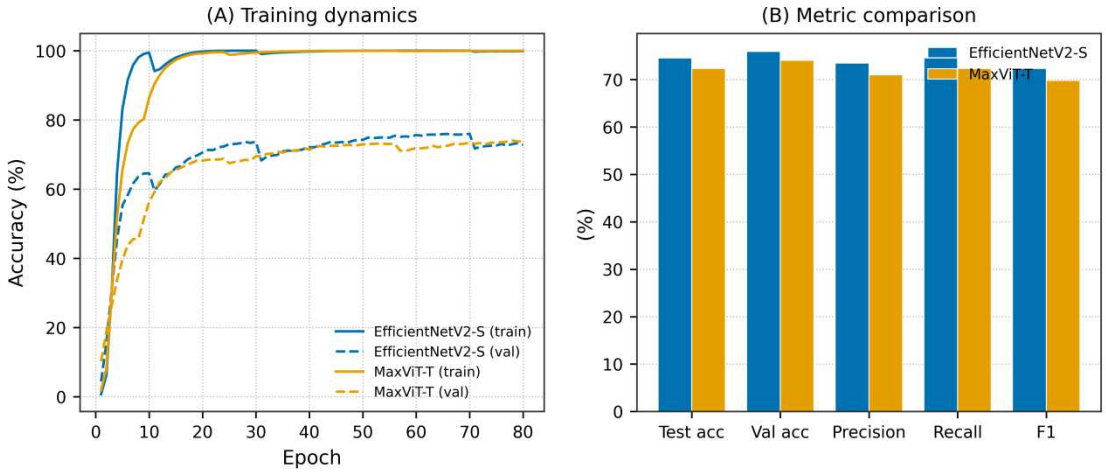


Fig. 3 Training dynamics of EfficientNetV2-S and MaxViT-T and comparison of five metrics

4.2 Analysis of Architectural Paradigms

Grouping the fourteen models by their five architectural families reveals a clear hierarchy of design vintages (Fig. 4). The two classical CNNs—WideResNet-50 (36.68%) and ResNeXt-50 (36.31%)—differ by only 0.37 pp despite a 2.7-fold parameter difference (69.73 M vs. 25.87 M), showing directly that the binding constraint within this family is not capacity. The lightweight CNNs (MobileNetV3-Large 51.21%, RepViT-M1.1 45.18%) differ by 6.03 pp and improve on the classical band by about 11.7 pp on average. The CNN–Transformer hybrids are the most heterogeneous family: MobileViT-V2-S (49.54%) falls far behind EfficientFormerV2-S (66.14%) and MaxViT-T (72.31%). A within-family range of 22.77 pp says that how local and global modelling are coupled matters far more than the hybrid label itself; it also exceeds by a factor of 7.7 the 2.95–2.97 pp separating the three design-contemporary family means. The modern ConvNets supply the accuracy frontier: EfficientNetV2-S (74.51%) is the global optimum, with EfficientNet-B3 (65.59%), InceptionNeXt-Small (62.51%), and ConvNeXt-V2-Tiny (59.86%) forming the body of the family (family mean 65.62%). The pure Transformers have the highest family mean, 68.59%: hierarchical designs lead (Swin Transformer 71.57%, PVTv2-B2 68.75%), while the isotropic ViT-B/16 (65.44%), despite being the largest and most computation-heavy network in the field, trails every hierarchically structured model. Family-mean accuracies span 32.09 percentage points in total; the statistical test of the family-level ordering is given below in this section.

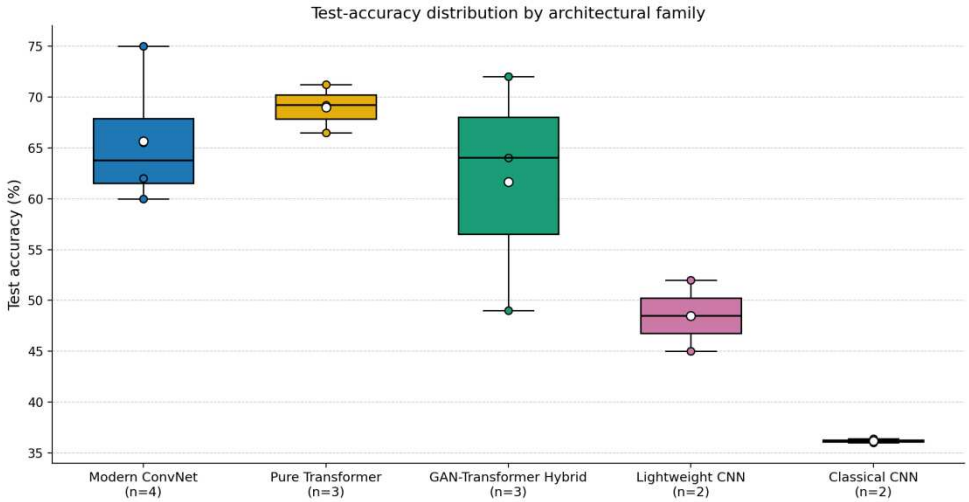


Fig. 4 Test-accuracy distributions by architectural family. Boxes show the interquartile range and median; points denote individual models. Family means span 32.09 percentage points.

How much does family membership actually explain? A descriptive one-way decomposition over the fourteen selected models attributes 79.9% of the variance in test accuracy to family ($F(4,9) = 8.96$, p = 0.003), but the result rests on the two extreme families: classical and lightweight CNNs supply only four of the fourteen models yet contribute 71.5% of the between-family sum of squares. Restricting the comparison to the three design-contemporary families that compete at the accuracy frontier (modern ConvNets, pure Transformers and CNN–Transformer hybrids, ten models in total) reverses the picture: family accounts for 11.2% of the variance ($F(2,7) = 0.44$, p = 0.66) and within-family differences for the remaining 88.8%, while the three family means are separated by only 2.95–2.97 pp against within-family ranges of 6.13–22.77 pp. An exact permutation test that reassigns the ten accuracies to family labels of identical sizes (4/3/3; 4,200 distinct partitions) returns p = 0.69 for the observed η^2 and p = 0.64 for the observed 5.92 pp spread of family means, so the ordering among these three families is indistinguishable from what random regrouping yields. The ordering is also sensitive to single members: the hybrid mean of 62.66% is depressed by MobileViT-V2-S (49.54%), the lightest model in the group at 6.49 M parameters, and removing that one member raises the hybrid mean to 69.22%, above both pure Transformers (68.59%) and modern ConvNets (65.62%), reversing the ordering; within the hybrid family, accuracy ranks in the same order as parameter count (32.51 M / 72.31%, 16.04 M / 66.14%, 6.49 M / 49.54%). These figures are computed over the fourteen models selected here and are not an inference to a wider population of architectures; they show that family membership behaves as a descriptive label co-varying with capacity and design vintage rather than as a manipulated factor. No ablation was run to separate macro-design from capacity, pre-training recipe or training schedule, so the family-level statements in this paper are associational.

Model size, by contrast, is essentially uninformative about accuracy (Fig. 5). Across the fourteen models, neither parameter count (Pearson $r = 0.00$, Spearman $\rho = +0.130$) nor FLOPs (Pearson $r = -0.034$, Spearman $\rho = +0.037$) carries any monotonic relationship with test accuracy. The starkest contrast is ViT-B/16. With 88.03 M parameters and 17.584 G MACs, the heaviest in the benchmark on both axes, it reaches only 65.44%, some 9.07 pp below EfficientNetV2-S while using four times its parameters and six times its computation. The penalty paid by small models runs in the opposite direction and is far steeper than a naive reading of the gaps suggests: the 6.49 M MobileViT-V2-S and the 5.31 M MobileNetV3-Large trail the best model by 24.97 and 23.30 pp respectively, so compressing a backbone by more than an order of magnitude in parameters costs close to a quarter of the accuracy range spanned by the entire field. Together with the

family analysis above, this yields the central empirical lesson of the benchmark: at 6,650-class fine-grained granularity with modern pre-training, accuracy covaries far more closely with the macro-design of an architecture than with its parameter budget. The observation is associational: macro-design was not manipulated independently of capacity, pre-training recipe or training schedule, and no ablation isolates its contribution.

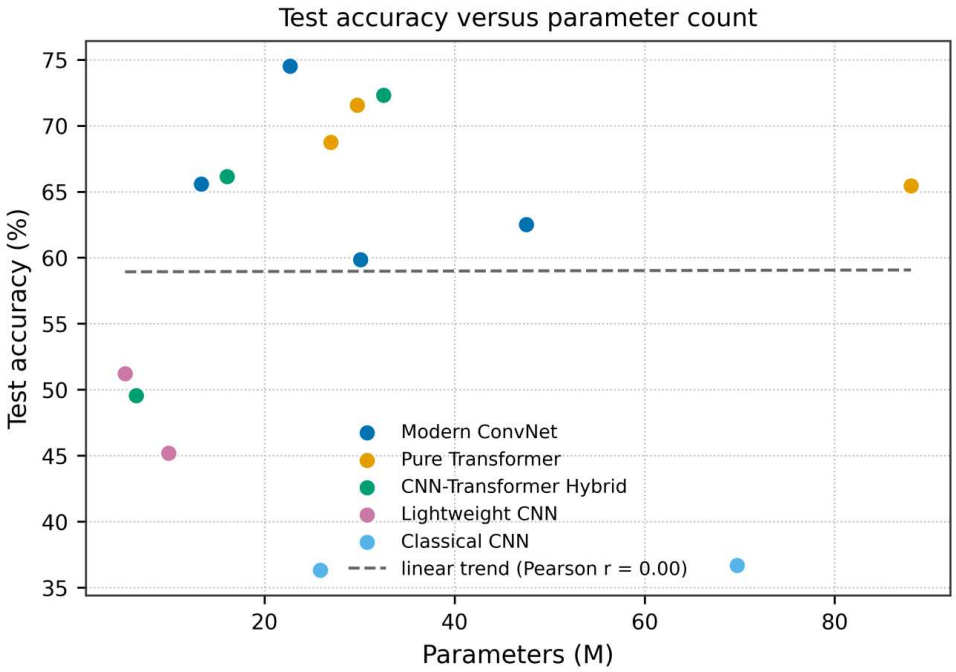


Fig. 5 Test accuracy versus parameter count, coloured by architectural family. The dashed line is the linear trend (Pearson r = 0.00): model size carries essentially no predictive power over accuracy.

4.3 Accuracy–Efficiency Trade-off

Benchmark evaluation concerns deployment cost as well as the accuracy ceiling. The behaviour of EfficientNetV2-S among the fourteen models carries clear implications for model selection: it attains the global accuracy optimum (74.51%) at markedly lower computational cost than the other leading models and reaches 66.23% validation accuracy by epoch 15, 87% of its final validation peak (Fig. 3), continuing the "small model, fast convergence" property observed on general image classification benchmarks [51]. For ancient script collation systems intended to run on ordinary office workstations, an architecture that reaches the accuracy ceiling at lower computational cost combines performance with deployability and thus carries particular engineering value. To characterise deployment properties completely, complexity and inference efficiency should be reported alongside accuracy: Table 6 gives the parameter count, computational cost, and single-sample inference latency of each model, and Fig. 6 plots accuracy against computational cost for all fourteen models so that the trade-off can be read off directly.

Table 6 Model complexity and inference efficiency.

Model	Parameters (M)	MACs (G)	Single-sample inference latency (ms)
EfficientNetV2-S	22.68	2.879	35.800
MaxViT-T	32.51	5.610	63.277
Swin Transformer	29.75	4.511	52.696
PVTv2-B2	26.95	3.905	41.503
EfficientFormerV2-S	16.04	1.275	34.502

Model	Parameters (M)	MACs (G)	Single-sample inference latency (ms)
EfficientNet-B3	13.33	0.995	26.7
ViT-B/16	88.03	17.584	63.743
InceptionNeXt-Small	47.53	8.362	66.293
ConvNeXt-V2-Tiny	30.10	4.472	34.455
MobileNetV3-Large	5.31	0.226	10.293
MobileViT-V2-S	6.49	1.438	23.709
RepViT-M1.1	9.88	1.359	23.306
WideResNet-50	69.73	11.428	98.520
ResNeXt-50	25.87	4.260	33.881

FLOPs are reported as multiply–accumulate operations (MACs) counted with fvcore, following the convention of [50], [51]. All fourteen models were re-measured in a single session under one identical protocol: the unified 6,650-way classification head ([512, 256], GELU), input $1 \times 3 \times 224 \times 224$, and the same tracer build—no figure is inherited from the earlier version of this benchmark. Because the attention operator aten::_native_multi_head_attention is not covered by the tracer, the ViT-B/16 figure includes an analytical supplement of 6.293 G for its twelve encoder layers (the supplemented total, 17.584 G, agrees with the commonly cited 17.568 G to within 0.1%). Latency was measured in the same session on a standalone CPU (Intel, GenuineIntel) / PyTorch 2.14.0+cpu / input $3 \times 224 \times 224$, batch = 1, model.eval() under torch.no_grad, median of 15 consecutive timings. Small deviations from commonly cited figures (e.g. MaxViT-T 5.610 vs. 5.383 G) arise from the heavier unified 6,650-way head and tracer coverage, not from re-evaluation of the backbones; absolute latencies should not be extrapolated to GPU hardware. The latency entry for EfficientNet-B3 is estimated rather than measured: that model entered the roster after the latency benchmarking session, so its value was obtained by transferring the median platform-scaling ratio of the other thirteen models (0.407) under an identical protocol; a regression against MACs gives 29.0 ms, indicating an uncertainty of about ±2 ms.

Table 6 and Fig. 6 show that no monotonic relationship holds between accuracy and computational cost. ViT-B/16 ranks first among all models in both parameter count (88.03 M) and computational cost (17.584 G), yet EfficientNetV2-S surpasses its 65.44% test accuracy with only 22.68 M parameters and 2.879 G, about a quarter of the parameters and a sixth of the computation. WideResNet-50 (11.428 G) and InceptionNeXt-Small (8.362 G) are the next most expensive, and EfficientFormerV2-S outscores them at 16.04 M parameters and 1.275 G.

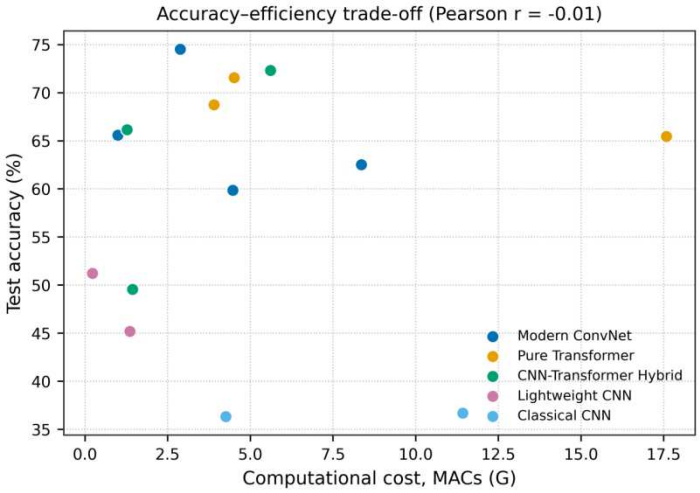


Fig. 6 Accuracy–efficiency trade-off (colour indicates architectural family)

Viewed as a Pareto frontier, the leading models define efficiency boundaries at different cost levels. EfficientNetV2-S achieves the highest accuracy at 2.879 G, whereas Swin Transformer (4.511 G), MaxViT-T (5.610 G), and InceptionNeXt-Small (8.362 G) require 1.6, 1.9, and 2.9 times its computation respectively without surpassing its accuracy. Accuracy leadership in this benchmark is therefore not bought at the highest computational price. Measured as accuracy returned per unit of computation, EfficientNet-B3 leads with 65.9 percentage points of test accuracy per GFLOP, reaching 65.59% with only 0.995 G and 26.7 ms; EfficientFormerV2-S follows (51.9 pp per GFLOP) and EfficientNetV2-S (25.9); PVTv2-B2 reaches 17.6 and MaxViT-T 12.9, whereas ViT-B/16 yields only 3.7, the least efficient by a wide margin. Under strict lightweight constraints, MobileNetV3-Large attains 51.21% at 0.226 G and 10.293 ms—the highest computation efficiency of the benchmark ($\approx$ 227 pp per GFLOP)—while MobileViT-V2-S offers 49.54% at only 6.49 M parameters, giving deployments a graded set of options along the frontier.

In inference latency, MobileNetV3-Large (10.293 ms), RepViT-M1.1 (23.306 ms), and MobileViT-V2-S (23.709 ms) form the low-latency band, yet all three concede more than 23 pp of accuracy to the optimum; EfficientNet-B3 sits just above this band at 26.7 ms while retaining 65.59%, which gives it the highest accuracy return per unit latency among the fast models. EfficientNetV2-S runs at 35.800 ms per sample, 3.5 times the lightest model yet roughly 57% and 54% of MaxViT-T (63.277 ms) and InceptionNeXt-Small (66.293 ms); within the leading group only ConvNeXt-V2-Tiny (34.455 ms) and EfficientFormerV2-S (34.502 ms) are faster, while the two largest pure Transformers pay a clear latency price (ViT-B/16 63.743 ms, Swin Transformer 52.696 ms). Weighing accuracy, computation, and latency together, EfficientNetV2-S remains the balanced choice for ancient script collation, whereas EfficientFormerV2-S is preferable when latency budgets are tighter and an 8.4 pp accuracy concession is acceptable.

The complexity and latency figures in Table 6 were measured in a single standalone CPU environment, whose framework version differs from the training environment listed in Table 4. Since parameter count and FLOPs are intrinsic architectural properties independent of the runtime, and latency is used only for cross-model comparison within the same environment, this setup does not affect the relative conclusions above, but absolute latency values should not be extrapolated to GPU deployment.

4.4 Error Structure and Long-Tail Effects

The value of a fine-grained benchmark lies in revealing the structural sources of error, not only in producing an overall ranking. Errors in fine-grained ancient character classification are expected to arise chiefly from three sources: high within-class variance caused by graphic variants, since the same character head varies greatly across four periods and multiple media; between-class confusion among visually similar heads that differ only at the component level; and the long-tailed distribution, under which tail categories lack training samples and are discriminated far less reliably than head categories. To quantify these effects, this paper reports a performance breakdown by category frequency band (Table 7, visualised in Fig. 7) and extracts the most typical confusion patterns from the confusion matrix (Table 8, visualised in Fig. 8).

Table 7 Performance by category frequency band

Category band	Categories	Test samples	Sample share (%)	Test accuracy (%)	Precision	Recall
Head (≥ 51)	155	2,956	23.49	90.32	0.9649	0.9032
Mid (16–50)	645	4,438	35.27	80.37	0.9008	0.8037
Tail (15)	5,850	5,189	41.24	60.44	0.7757	0.6044

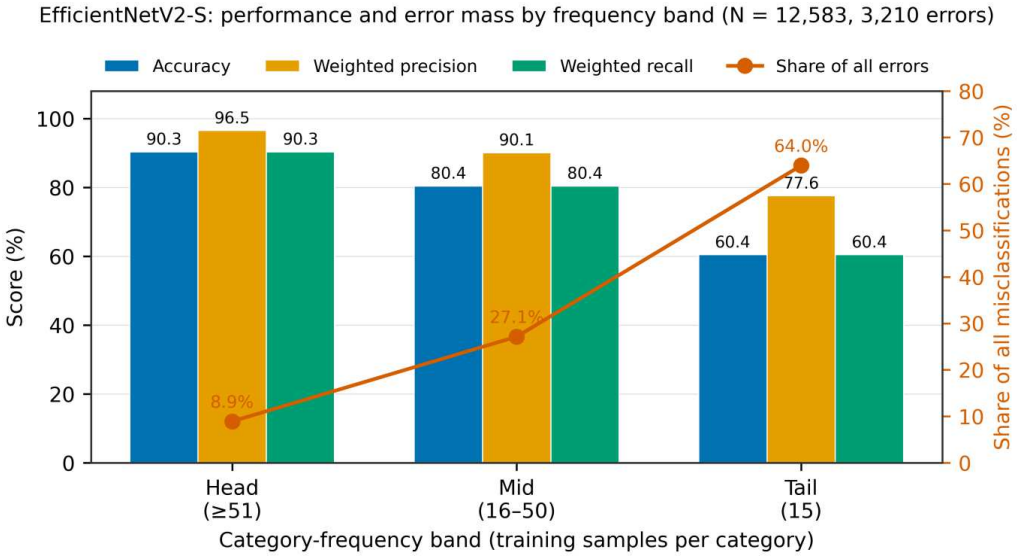


Fig. 7 Performance decomposition by character-head frequency band under EfficientNetV2-S

Grouped bars report accuracy, weighted precision, and weighted recall for the head (≥ 51), mid (16–50), and tail (15) bands; the overlaid line (right axis) gives each band's share of all 3,210 misclassifications.

Table 7 presents the performance breakdown over three bands defined by the per-category training sample count (after the dataset's balanced offline augmentation, which gives every category at least 15 samples; the tail band thus comprises categories with no source material beyond this baseline, and the head threshold is raised to 51 accordingly). Best model: EfficientNetV2-S. The breakdown was computed on the per-sample predictions of the best model under the full 6,650-category setting and uses exactly the same evaluation convention as Table 5 (the sample-weighted average of the three band accuracies is 74.49%, differing from the 74.51% reported in Table 5 by 0.02 percentage points, a checkpoint-serialisation difference of three samples that does not affect any conclusion here). The results confirm a long-tail effect with a steep monotonic gradient: test accuracy falls from 90.32% for head categories to 80.37% for mid and 60.44% for tail, a head-to-tail drop of 29.88 percentage points. More diagnostic is the structure of error distribution. Tail

categories contribute 64.0% of all errors (2,053 of 3,210) while holding 41.24% of test samples, an error density 1.55 times their sample share; head categories behave in the opposite way, contributing 8.9% of errors (286 of 3,210) from 23.49% of samples, a factor of 0.38. The long tail therefore does not raise the overall error rate uniformly but concentrates error onto sample-scarce categories. Comparing the sample-weighted and category-equal conventions, the tail band has weighted precision and recall of 0.7757 and 0.6044, whereas its macro-averaged recall is 0.5653, a gap of about 3.9 percentage points; macro-averaged recall is 0.8982 and 0.7942 for the head and mid bands respectively. This discrepancy shows that performance varies more widely across individual tail categories — 940 of the 2,647 evaluated tail categories fall below 50% recall, 604 of them represented by a single test sample — and that the collapse of a few extremely rare categories is smoothed away by sample weighting. It should also be noted that 3,203 of the 6,650 categories (48.2%) have no samples in the test set and all belong to the tail band, so that 2,647 tail categories actually participate in evaluation and the test set covers 3,447 categories in total; the discriminative performance of these zero-sample categories cannot be determined by this benchmark, which defines a clear boundary on evaluation coverage. Combining this with the dataset structure (tail categories amount to 88.0% of all categories and, after uniform augmentation to the 15-sample baseline, still hold 74.1% of the training images, whereas the model attains only 60.44% accuracy on their test samples), the naive mitigation strategy of offline augmentation to a uniform baseline is far from sufficient under the present configuration: further gains in overall accuracy should come from dedicated long-tail recognition methods rather than from simply enlarging the sample set. Fig. 7 visualises the band-level accuracy and error-concentration contrast.

Table 8 Top-10 most confused character-head pairs of the best model, EfficientNetV2-S

No.	True character head	Predicted as	Errors	Test samples of this head	Within-head error rate (%)
1	匕 (bi)	人 (ren)	5	7	71.4
2	祿 (lu)	录 (lu)	4	5	80.0
3	玉 (yu)	王 (wang)	4	11	36.4
4	祖 (zu)	且 (qie)	3	8	37.5
5	聽 (ting)	聖 (sheng)	3	4	75.0
6	臽 (xian)	旨 (zhi)	3	3	100.0
7	鬳 (yan)	鼎 (ding)	3	5	60.0
8	丞 (cheng)	兵 (bing)	3	6	50.0
9	末 (mo)	木 (mu)	3	5	60.0
10	冢 (meng)	家 (jia)	3	4	75.0

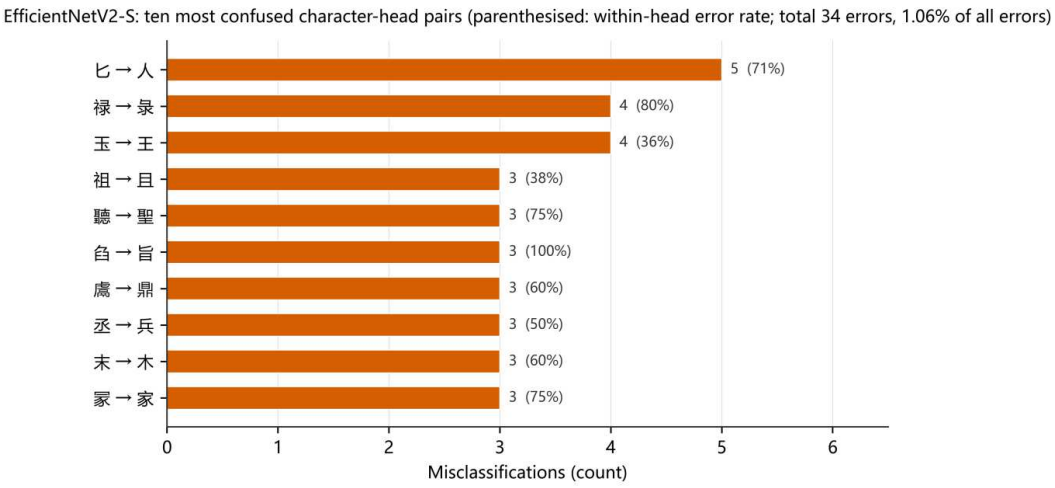


Fig. 8 Top-10 most confused character-head pairs of EfficientNetV2-S.

Bar length is the number of misclassified samples; the label reports the error count together with the within-head error rate (errors of that head divided by its test samples). All ten pairs involve different character heads.

Table 8 lists the ten character-head pairs with the highest error counts. Every one of the 3,210 errors crosses character heads; none falls inside a head's own variants. The base for that statistic is itself worth stating: only 34 character heads have two or more of their classes represented in the test set, covering 111 test samples (0.88%) in which same-head confusion is possible at all. Of the 49 errors falling inside this subset, none is a same-head confusion. The same pattern holds across all twelve analysed models (at most three same-head errors each, never exceeding 0.06% of a model's errors). Errors therefore arise almost entirely from visual similarity between different character heads ("yu" 玉 confused with "wang" 王, "mo" 末 with "mu" 木, "bi" 匕 with "ren" 人). In other words, the fine-grained difficulty posed by OBWX lies in distinguishing graphically similar characters rather than in modelling variation within a single head, a pattern consistent with annotation that stays coherent across periods. Examining the top-ten pairs shows that they are concentrated on sample-scarce heads (3–11 test samples each) with within-head error rates between 36.4% and 100.0%, yet the ten pairs together account for only 34 errors, 1.06% of the total — a highly dispersed error distribution with no dominant confusion pattern. Two ways forward follow from this pattern: dedicated long-tail recognition for the scarce heads, and targeted enhancement of similar-glyph discrimination through component-level attention or part decoupling. The second should pay off more directly than generic resampling, because the errors sit on individually rare heads rather than piling onto a single confusion. Fig. 8 plots these ten pairs together with their error counts and within-head error rates.

4.5 Statistical Significance and Sampling Uncertainty

To determine whether the observed ranking reflects real differences rather than a single-run artefact, we report 95% Wilson confidence intervals for every model computed over the 12,583 test samples (Table 9, Fig. 9). The best model, EfficientNetV2-S (74.51%, 95% CI [73.74, 75.27]), is statistically separated from every other model: its interval lies entirely above that of the runner-up MaxViT-T (72.31%, [71.52, 73.09]). MaxViT-T and Swin Transformer ([70.78, 72.35]) overlap, as do the EfficientFormerV2-S–EfficientNet-B3–ViT-B/16 trio ([65.31, 66.96], [64.75, 66.41], [64.60, 66.26]) and the MobileNetV3-Large–MobileViT-V2-S pair ([50.34, 52.08] and [48.66, 50.41]); the fine-grained ordering within these groups should be read as indicative rather than decisive, while all remaining adjacent pairs separate cleanly. The two classical CNNs

([35.84, 37.52] and [35.47, 37.16]) overlap with each other but are statistically distinct from every other model, so the family-level gaps identified in Section 4.2 are real rather than sampling noise. Run-to-run stability is examined in Section 4.7, where the full fourteen-backbone benchmark is replicated under three random seeds (42, 123, 2024). Because every model is evaluated on the identical fixed test set, these intervals answer a well-defined question under the unified protocol: which accuracy differences are real. These intervals assume independent samples. Multiple glyphs of the same character head are correlated, so the effective sample size is smaller and the intervals approximate. No multiplicity correction is applied across the 91 pairwise comparisons.

Table 9 Test accuracy (%) of the fourteen models on the 12,583-sample OBWX test set with 95% Wilson confidence intervals ($N = 12{,}583$)

Model	Test accuracy (%)	95% CI (%)
EfficientNetV2-S	74.51	[73.74, 75.27]
MaxViT-T	72.31	[71.52, 73.09]
Swin Transformer	71.57	[70.78, 72.35]
PVTv2-B2	68.75	[67.94, 69.56]
EfficientFormerV2-S	66.14	[65.31, 66.96]
EfficientNet-B3	65.59	[64.75, 66.41]
ViT-B/16	65.44	[64.60, 66.26]
InceptionNeXt-Small	62.51	[61.66, 63.35]
ConvNeXt-V2-Tiny	59.86	[59.00, 60.71]
MobileNetV3-Large	51.21	[50.34, 52.08]
MobileViT-V2-S	49.54	[48.66, 50.41]
RepViT-M1.1	45.18	[44.31, 46.05]
WideResNet-50	36.68	[35.84, 37.52]
ResNeXt-50	36.31	[35.47, 37.16]

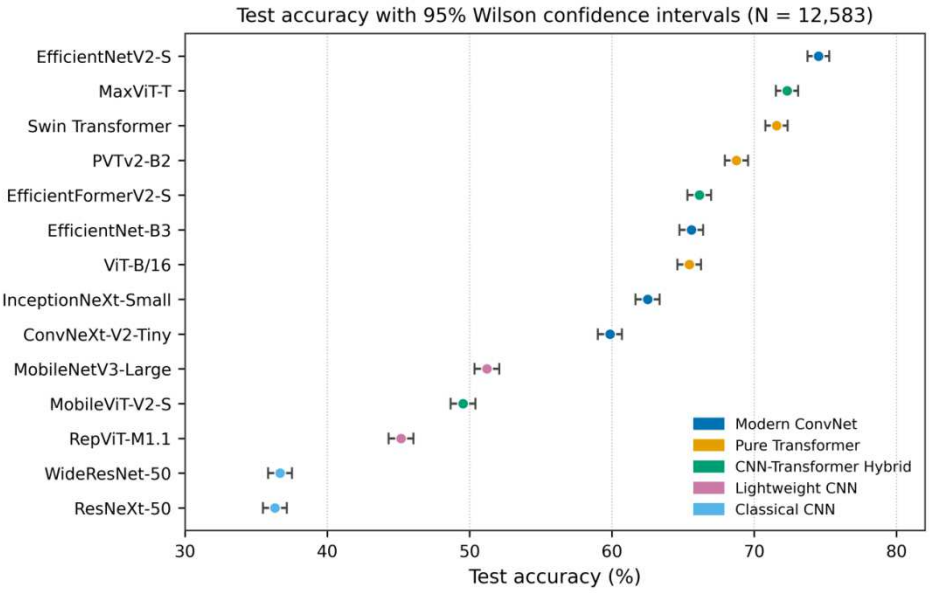


Fig. 9 Per-model test accuracy with 95% Wilson confidence intervals ($N = 12{,}583$). Points show the point estimates, horizontal lines the 95% confidence intervals. Non-overlapping intervals (e.g. the best versus the worst models) indicate statistically significant differences.

4.6 Robustness and Failure-Mode Analysis

The error structure in Section 4.4 points to two robustness stresses, both already measurable with the data reported above. First, the long tail: 3,203 of 6,650 categories (48.2%) have no test sample and all belong to the tail band, while 940 of the 2,647 evaluated tail categories fall below 50% recall; rare characters are

therefore recognised unreliably, and any realistic deployment will meet glyphs whose category was never represented at test time. Second, fine-grained visual confusion: the top-10 error pairs all occur between different character heads of high component similarity (玉→王, 末→木), and the error mass is highly dispersed (the ten pairs account for only 1.06% of all errors), so no single confusion dominates. A third stress, however, cannot be measured from clean-set results alone: robustness to the stronger degradations found in authentic rubbings: stroke loss, ink bleed, plate noise, and OCR-style artefacts. Training augmentation was deliberately restricted to conservative, fidelity-preserving transforms (±5° rotation, ±10% brightness/contrast, no horizontal flipping) so as not to manufacture invalid glyphs, and the test split is clean by construction. To quantify this third stress we apply a controlled noise-injection protocol to the frozen clean test split. Three perturbation families are involved: additive Gaussian noise ($\sigma \in \{15, 30, 50\}$ on the 0–255 pixel scale), salt-and-pepper impulse noise (fraction $p \in \{0.03, 0.07, 0.12\}$), and stroke erosion to simulate stroke loss and ink bleed (structuring element $k \in \{2, 3, 5\}$). All act on the uint8 pixels before normalisation, and every model sees identical perturbation patterns, so differences between models cannot be attributed to sampling variance. We run it on the three leading backbones of Table 5, EfficientNetV2-S, MaxViT-T and Swin Transformer, which between them cover the accuracy frontier and three architectural families; their clean-set accuracies re-measured in this pass agree with Table 5 to within 0.04 pp, far below the half-width of the Wilson intervals of Section 4.5. Table 10 reports the results and Fig. 10 visualises them.

Three findings emerge. (i) Degradation is monotone in perturbation strength for every model and every family, and even the mildest perturbation (Gaussian $\sigma = 15$) costs 2.86–8.93 pp, more than the 2.21 pp separating the top two models on clean data, so differences observed there stop being resolvable once inputs degrade. (ii) The clean-set ordering does not survive perturbation. MaxViT-T, second on clean accuracy (2.21 pp behind EfficientNetV2-S in this pass, 2.20 pp in Table 5), is the most robust of the three: it loses on average 12.30 pp across the nine perturbation settings (retaining 83.0% of its clean accuracy), is the best of the three in six of the nine settings, and is never the worst; its 2.21 pp clean-set deficit turns into a 6.31 pp advantage under Gaussian noise at $\sigma = 50$. EfficientNetV2-S, the clean-set winner, loses 16.82 pp on average (77.4% retained) and ranks last of the three in four settings, including all three Gaussian levels. Swin Transformer loses 22.55 pp on average (68.5% retained) and is the worst in five settings, although on the clean set it trails MaxViT-T by only 0.75 pp. (iii) The failure mode is architecture-specific. Stroke erosion is the perturbation closest to authentic rubbing damage, and it is catastrophic for Swin Transformer: it already costs 35.84 pp at $k = 2$ (49.9% of clean accuracy retained) and leaves the model at 30.02% at $k = 5$, whereas the same perturbation costs EfficientNetV2-S 14.87 pp and MaxViT-T 12.60 pp. Averaged over the three models, salt-and-pepper and stroke erosion are comparably damaging at the strongest level (−25.18 and −28.21 pp) while Gaussian noise is markedly milder (−13.68 pp). Perturbation also widens the spread between the three models from 2.96 pp on the clean set to 23.99 pp under erosion at $k = 2$, roughly eight-fold. Degradation amplifies architectural differences rather than washing them out.

Two scope conditions bound this evidence. First, the sweep covers three of the fourteen backbones; the remaining eleven will follow under the identical protocol. The three reported here are the accuracy leaders rather than a family-stratified sample, so how the lighter models rank on robustness is not yet known, and they may well respond differently to impulse noise. Second, synthetic perturbations are proxies: they control severity precisely but do not reproduce the statistical structure of real rubbing degradations, and validation on authentically degraded scans remains future work. Sections 5.3 and 6 return to both points.

Table 10 Noise robustness of the three leading backbones (%). Test accuracy on the frozen clean test split ($N = 12{,}583$) under three perturbation families at three severity levels each; the value in parentheses is the

change relative to the same model's clean-set accuracy, in percentage points. All models receive identical perturbation patterns (shared random seed, seed 42); noise is applied to the uint8 pixels before normalisation. Δ is the mean change over the nine perturbation settings; retention is mean noisy accuracy divided by clean accuracy.

Perturbation family	Level	EfficientNetV2-S	MaxViT-T	Swin Transformer
Clean (reference)	—	74.49	72.28	71.53
Additive Gaussian noise	$\sigma = 15$	65.56 (−8.93)	69.42 (−2.86)	65.59 (−5.94)
	$\sigma = 30$	60.76 (−13.73)	65.81 (−6.47)	61.89 (−9.64)
	$\sigma = 50$	56.62 (−17.87)	62.93 (−9.35)	57.72 (−13.81)
Salt-and-pepper noise	p = 0.03	60.22 (−14.27)	62.15 (−10.13)	63.93 (−7.60)
	p = 0.07	55.07 (−19.42)	57.24 (−15.04)	51.74 (−19.79)
	p = 0.12	50.81 (−23.68)	52.56 (−19.72)	39.40 (−32.13)
Stroke erosion	$k = 2$	59.62 (−14.87)	59.68 (−12.60)	35.69 (−35.84)
	k = 3	58.42 (−16.07)	58.39 (−13.89)	34.82 (−36.71)
	$k = 5$	51.99 (−22.50)	51.66 (−20.62)	30.02 (−41.51)
Mean over nine settings	—	57.67 (−16.82)	59.98 (−12.30)	48.98 (−22.55)
Retention of clean accuracy (%)	—	77.4	83.0	68.5

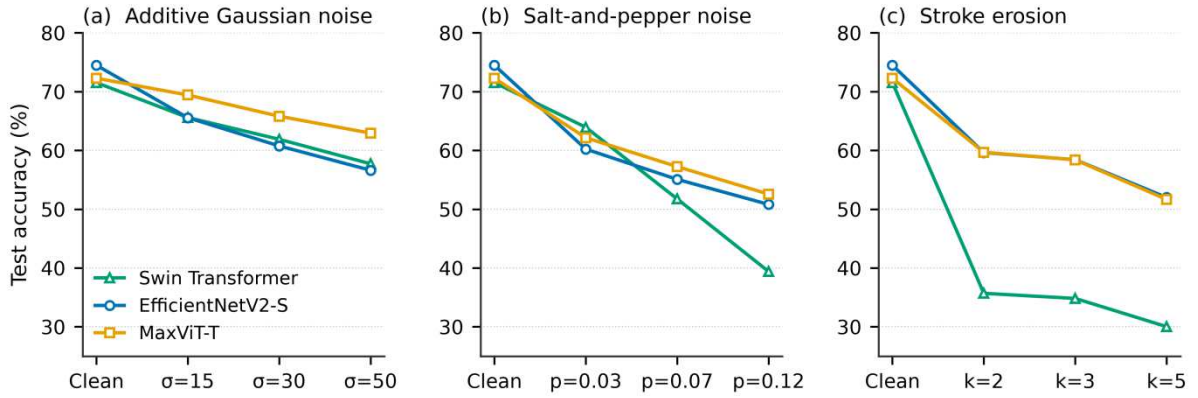


Fig. 10 Test accuracy of the three leading backbones under controlled perturbations. (a) Additive Gaussian noise; (b) salt-and-pepper impulse noise; (c) stroke erosion. The leftmost point of each curve is the clean-set accuracy; all three panels share the same vertical scale, so slopes are directly comparable across panels.

4.7 Multi-Seed Stability and Variance Analysis

Sections 4.1–4.6 rest on a single training run per backbone. The Wilson intervals reported in Section 4.5 quantify only the sampling uncertainty of the fixed 12,583-image test set; they do not cover the randomness inherent in training itself: parameter initialisation, mini-batch ordering, and the stochastic augmentation draw. A ranking observed under one seed could therefore still reflect run-to-run fluctuation rather than a genuine architectural difference. To close this gap, the complete benchmark was replicated under two additional random seeds, 123 and 2024: each of the fourteen backbones is trained three times under the identical protocol of Table 3: the same stratified partition, the same fidelity-constrained augmentation, the same 80-epoch budget, and the same unified 6,650-way MLP head. Only the random seed is changed; no hyper-parameter is tuned for any seed or any model. The seed-42 results are those reported in Table 5.

For every backbone we report the test accuracy of each seed, the three-seed mean with its sample standard deviation, and the seed spread, defined as the difference between the highest and the lowest test accuracy over the three runs. The spread is the quantity of interest: it bounds the displacement that a single-run ranking could suffer from training randomness alone. A position in Table 5 is treated as stable when the seed spread is smaller than the accuracy gap to its nearest neighbour in the ranking, and as tied when the

spread exceeds that gap.

Table 11 reports the completed three-seed results, and Fig. 11 visualises them. Eleven of the fourteen backbones are stable at the resolution this replication can detect: their spreads range from 0.15 pp (EfficientNet-B3) to 2.60 pp (PVTv2-B2), with a median of 0.57 pp. Judged by that criterion, eight of the fourteen positions in Table 5 are stable; the remaining six involve adjacent gaps of the same order as the seed noise (for instance the EfficientNet-B3 / ViT-B/16 pair, whose 0.15 pp gap should be read as tied). The head of the ranking is unaffected: EfficientNetV2-S ranks first under all three seeds (74.51 / 74.25 / 74.60), and the set of the three leading backbones {EfficientNetV2-S, MaxViT-T, Swin Transformer} is identical in every run, MaxViT-T and Swin swapping the second and third positions within 0.91 pp. Rank correlations between seeds are correspondingly high (Spearman $\rho = 0.88 - 0.96$).

Three backbones violate this pattern outright: ConvNeXt-V2-Tiny (spread 13.72 pp), ResNeXt-50 (15.25 pp), and WideResNet-50 (24.75 pp). For the latter two, the seed-42 run diverges from the other two seeds from the earliest epochs (validation accuracy at epoch 10: 21.18% versus 51.79% and 51.13% for WideResNet-50; 19.60% versus 37.72% and 38.30% for ResNeXt-50), whereas seeds 123 and 2024 agree closely with each other throughout. ConvNeXt-V2-Tiny behaves the other way around: 59.86 / 53.46 / 46.13, with no two seeds agreeing early in training. For this architecture, optimisation itself appears unstable at the fixed budget. Both effects are confined to the lower half of Table 5: all three backbones lie at least 14.6 pp below the leader there, so the head-of-ranking conclusions of Sections 4.1–4.5 are unaffected. At the family level, the three-seed means preserve the ordering of the three upper families but swap the two lowest (classical 49.61% versus lightweight 48.60% on three-seed means, against 36.50% versus 48.20% under seed 42)—a further consequence of the anomalous seed-42 runs of the two classical models.

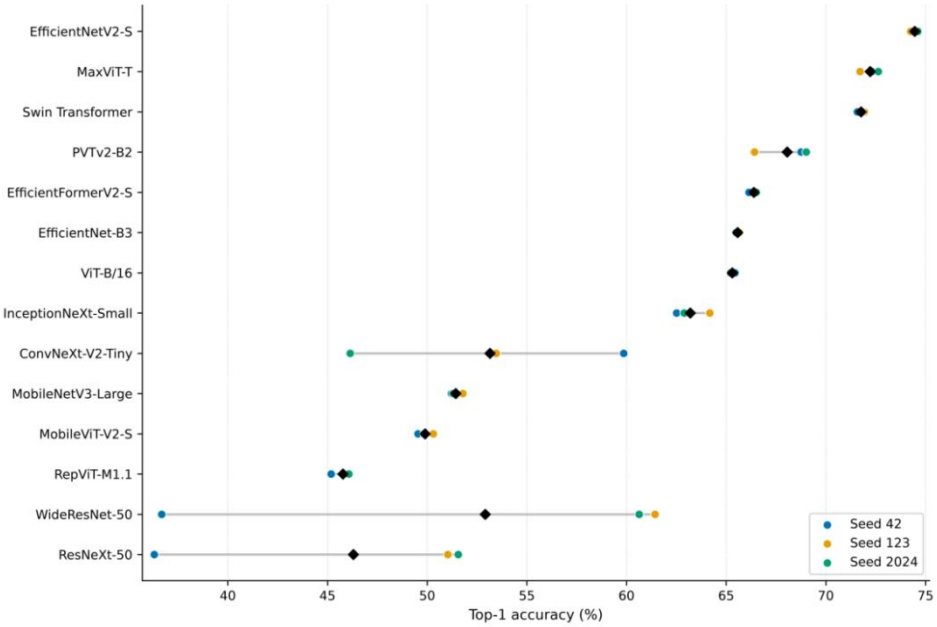


Fig. 11 Three-seed replication of all fourteen backbones under the identical protocol of Table 3 (seeds 42, 123, 2024). Seed 42 reproduces Table 5; seeds 123 and 2024 are independent full 80-epoch training runs. Models are ordered by seed-42 accuracy; only the random seed differs, with no hyper-parameter tuned for any seed or model.

Table 11 Test accuracy (%) of the fourteen models under three random seeds on the OBWX test set. Seed 42 reproduces Table 5; seeds 123 and 2024 are independent full 80-epoch runs made under that same protocol. SD is the sample standard deviation over the three runs and Spread is the difference between the highest and the lowest accuracy.

Model	Seed 42	Seed 123	Seed 2024	Mean ± SD (%)	Spread (pp)
EfficientNetV2-S	74.51	74.25	74.60	74.45 ± 0.18	0.35
MaxViT-T	72.31	71.72	72.63	72.22 ± 0.46	0.91
Swin Transformer	71.57	71.92	71.80	71.76 ± 0.18	0.35
PVTv2-B2	68.75	66.42	69.01	68.06 ± 1.43	2.60
EfficientFormerV2-S	66.14	66.50	66.50	66.38 ± 0.21	0.37
EfficientNet-B3	65.59	65.64	65.49	65.58 ± 0.08	0.15
ViT-B/16	65.44	65.24	65.22	65.30 ± 0.12	0.21
InceptionNeXt-Small	62.51	64.17	62.89	63.19 ± 0.87	1.66
ConvNeXt-V2-Tiny	59.86	53.46	46.13	53.15 ± 6.87	13.72
MobileNetV3-Large	51.21	51.78	51.31	51.43 ± 0.31	0.57
MobileViT-V2-S	49.54	50.31	49.85	49.90 ± 0.39	0.77
RepViT-M1.1	45.18	46.06	46.07	45.77 ± 0.51	0.89
WideResNet-50	36.68	61.42	60.63	52.91 ± 14.06	24.75
ResNeXt-50	36.31	51.03	51.56	46.30 ± 8.66	15.25

5 Discussion

5.1 Comparison with Existing Studies

In terms of task positioning, this benchmark complements rather than replaces existing evaluations of ancient script recognition. The HUST-OBC baseline attains 94.6% accuracy and 0.914 macro-averaged F1 on 1,588 deciphered oracle bone character heads [4], fewer than a quarter of the categories here and covering only the single period of oracle bone script. OBC306 [18] is known for 309,551 rubbing images but has only 306 categories with a highly imbalanced distribution. The graph-structured evolution dataset [22] pioneered evolution-chain modelling, yet 756 chains cannot support large-scale classification evaluation. No existing resource offers all three at once: cross-period coverage, a large category space, and fine granularity. At 6,650 categories drawn from authoritative sources and spanning four consecutive periods, this benchmark fills that gap. Numbers do not transfer between benchmarks, but the pattern they share is clear enough. Given sufficient inter-category difference, convolutional networks reach a high discriminative ceiling on ancient script images, as the HUST-OBC result above illustrates. Oracle-bone-specific recognisers such as Prism-OBI [38], which decouples visual perception from semantic interpretation and applies structural denoising, optimise for the single-period oracle bone setting under divergent splits and augmentations; their reported accuracies are therefore not directly comparable to OBWX, whose contribution is precisely the standardised cross-period protocol that makes such comparisons possible in the first place.

Our finding that convolutional and hybrid architectures beat pure transformers agrees with a common observation in recent ancient-script studies, namely that CNN inductive bias helps [16], [40], [41]. EfficientNetV2-S taking first place, with Swin ahead of ViT, sharpens the point further. The results are consistent with locality priors mattering: a convolutional architecture distinguished by efficient local modelling (EfficientNetV2-S) attains the global optimum, and a hybrid design that fuses local and global information (MaxViT-T) follows 2.20 pp behind. This provides direct guidance for backbone selection in ancient character recognition systems.

Regarding long-tail handling, this benchmark adopts the naive mitigation of offline augmentation of tail categories to 15 images and does not incorporate dedicated long-tail methods such as decoupled representation learning [10] or diffusion-based pseudo-sample generation [40]. The results therefore reflect the “bare performance” of each backbone under a single naive protocol. That gives later work on long-tail methods a clean control baseline, and the direction is a natural one to take next.

5.2 Usability and Reproducibility

Usability has two sides, data and protocol. On the data side, three independent checks point the same way, though none of them proves that the labels are sound. First, model selection by validation accuracy transfers cleanly to the test set: ranking the fourteen backbones by best validation accuracy reproduces the test ranking at 12 of the 14 positions, and the single exception (EfficientNet-B3 versus ViT-B/16) is a 0.15 pp noise-level swap. Second, across the twelve models for which per-sample predictions are available, same-head confusion is absent for the best model and bounded at three errors for any single model, never exceeding 0.06% of that model's errors, indicating that cross-period variants of one character head barely bleed into one another. Third, accuracy on the head band—categories with at least 51 training samples—reaches 90.32%, consistent with a category space that is learnable once samples suffice. OBWX is also released on Zenodo under a CC BY 4.0 licence, with traceable provenance and barrier-free access [6]. At the protocol level, the benchmark publishes the complete partitioning scheme, augmentation policy, training configuration, and evaluation metrics, and the training and evaluation code is open-sourced under the MIT licence. With the random seed fixed, any later study can reproduce these results under exactly the same conditions and compare new methods against them. Together, these two sides are what make the benchmark worth keeping as a public reference.

5.3 Limitations

Six limitations apply. First, run-to-run stability rests on three seeds, enough to expose gross instability but too few to pin down the variance of individual positions. The replication of Section 4.7 confirms the head of the ranking: eleven of fourteen backbones have spreads of 0.15–2.60 pp (median 0.57 pp), and the leading model and the top-three set are identical in every run. It also reveals that ConvNeXt-V2-Tiny, WideResNet-50, and ResNeXt-50 have seed spreads of 13.7–24.8 pp, with the seed-42 runs of the two classical models diverging from the other two seeds from the earliest epochs. The cause could not be established after the fact, so the seed-42 entries of these models in Table 5 carry that caveat. Second, long-tail handling uses a naive augmentation strategy and does not incorporate dedicated methods such as decoupled representation learning [10] or diffusion-based pseudo-sample generation [40]; the benchmark therefore reflects “bare performance” under a single naive protocol, which nevertheless offers a clean control baseline for subsequent long-tail research. Third, error characterisation provides the long-tail band breakdown and the analysis of confusing character-head pairs (Tables 7 and 8) and discloses the coverage boundary of the test set at 3,447 of 6,650 categories, but the per-period breakdown (oracle bone / bronze / Warring States / small seal) has not yet been completed: the processed data are organised by character head and do not retain per-sample period labels, which must be added in a later version. Fourth, robustness to degraded inputs is quantified only for the three leading backbones: Section 4.6 reports a controlled noise-injection sweep on EfficientNetV2-S, MaxViT-T, and Swin Transformer and shows that accuracy rankings obtained on clean inputs do not carry over to degraded ones, but the remaining eleven backbones and validation against authentic rubbing artefacts (stroke loss, ink bleed, plate noise) remain to be completed, so the robustness conclusions cannot yet be extended to

lighter architectures. Fifth, architectural family is an observational label rather than a manipulated factor: Section 4.2 shows that among the three design-contemporary families, family membership accounts for only 11.2% of the variance in test accuracy (permutation $p = 0.69$), that removing the single lightest member raises the hybrid mean from 62.66% to 69.22% and reverses the family ordering, and that no ablation separates macro-design from capacity, pre-training recipe or training schedule; every family-level statement in this paper is therefore associational. Sixth, the evaluated models are predominantly general-purpose image classification backbones, and the benchmark does not cover sequence modelling methods specific to text recognition or recent large multimodal models. These limitations at the same time define concrete directions for future work.

6 Conclusion and Future Work

On the OBWX dataset (74,047 images, 6,650 ideographic character categories, four historical periods) we have built one common reference for cross-period fine-grained classification of ancient Chinese characters. One protocol governs everything: stratified partitioning, fidelity-constrained augmentation, long-tail mitigation, and sample-weighted metrics. Fourteen representative architectures spanning five families were then trained and tested under strictly identical conditions. EfficientNetV2-S achieves the best result at 74.51% test accuracy, followed by MaxViT-T (72.31%) and the hierarchical Swin Transformer (71.57%), and is the only model statistically separated from the rest. Between family means the spread is 32.09 percentage points, running from the classical-CNN mean of 36.50% to the pure-Transformer mean of 68.59%. Neither parameter count nor FLOPs correlates with accuracy ($|r| \leq 0.04$). At this granularity, accuracy tracks design vintage far more closely than it tracks capacity, and the 38.20 pp best-to-worst span shows how hard a 6,650-category space is to separate. Within the pure-Transformer family, hierarchical designs (Swin, PVTv2) clearly outperform the isotropic ViT-B/16, localising the earlier "Transformers lag" observation to architecture-internal structure rather than to attention itself. Measured against 95% confidence intervals over the 12,583-sample test set, the family gaps exceed sampling fluctuation. The three-seed replication resolves the bottom of the ranking: the ordering of the three upper families is preserved, whereas the two lowest families swap because the seed-42 runs of the two classical CNNs are anomalous. Under seed 42 the classical-CNN band falls below 37%, evidence that the benchmark separates architectural designs sharply.

In latency, ConvNeXt-V2-Tiny (34.455 ms), EfficientFormerV2-S (34.502 ms), and EfficientNetV2-S (35.800 ms) are the fastest of the leading group, at 54–57% of the latency of MaxViT-T (63.277 ms), while EfficientNet-B3 reaches 26.7 ms with 65.59% accuracy.

Five directions follow. First, long-tail methods such as difficulty-aware decoupled learning [10] and diffusion-based pseudo-sample generation [40] will be introduced onto the present "bare performance" baseline to quantify the gain of dedicated strategies over naive augmentation. Second, the 855 characters with complete four-period sequences in OBWX will support research on evolution trajectory prediction and temporal modelling [6], extending character-head classification into explicit analysis of graphic evolution. Third, a per-period performance breakdown will be added, and the similar-glyph confusion revealed in Table 8 will be addressed by combining local attention with component-level modelling [39] to improve discrimination of confusing pairs. Fourth, drawing on progress in OBI-Bench [11] and V-Oracle [12], the joint evaluation of classification benchmarks with the reasoning capabilities of large multimodal models will be explored, advancing ancient character recognition from isolated tasks toward an integrated scholarly ecosystem [3]. Fifth, extend the noise-injection protocol of Section 4.6 from the three leading backbones to

all fourteen, and, going beyond synthetic perturbations, validate robustness against authentic rubbing degradations: lost strokes, ink bleed, plate noise, and OCR artefacts. That would quantify what these results lose in real deployment, and in particular whether the fragility of hierarchical attention under stroke erosion extends beyond Swin Transformer.

Declarations

Author contributions: Yanqiang Ge: Conceptualization, Methodology, Supervision, Project administration, Writing – review & editing. Zixu Yang: Software, Investigation, Data curation, Writing – original draft. Ziqi Ge: Formal analysis, Data curation. Zhan Zhang: Software, Validation, Resources. Haoyu Wang: Investigation, Resources.

Competing interests: All authors declare no financial or non-financial competing interests.

Data availability: The OBWX dataset (v1.0) used in the benchmark experiments is released on Zenodo under a CC BY 4.0 licence (DOI: 10.5281/zenodo.20037584) [6]. The training and evaluation code is open-sourced under the MIT licence.All benchmark source code, evaluation pipelines and experimental logs are available at https://github.com/yangzixu-000/code.

Ethics statement: This study involves neither human participants nor animal experimentation, and the OBWX dataset is released under a CC BY 4.0 licence; no ethical approval was required.

Acknowledgements: The authors thank the Key Laboratory of Oracle Bone Inscriptions Information Processing (Anyang Normal University), Ministry of Education, for its support of the expert verification of the OBWX dataset.

Funding： This research was funded by the Henan Province Teacher Education Curriculum Reform Research Project, grant number 2026-JSJYYB-036, and the Henan Province Higher Education Teaching Reform Research and Practice Project, grant number 2026SJGLX150.

References

[1] Qiu, X. *Chinese Writing* (Commercial Press, 2013). (in Chinese).

[2] Liu, Z. *Structure of Ancient Chinese Characters* (Fujian People's Publishing House, 2006). (in Chinese).

[3] Ma, Y. R. Towards computational Chinese paleography. Preprint at arXiv https://arxiv.org/abs/2601.06753 (2026).

[4] Wang, P. J., Zhang, K. L., Wang, X. Y. et al. An open dataset for oracle bone character recognition and decipherment. *Sci. Data* **11,** 976; 10.1038/s41597-024-03807-x (2024).

[5] Guan, H. S., Wan, J. P., Liu, Y. L. et al. An open dataset for the evolution of oracle bone characters: EVOBC. Preprint at arXiv https://arxiv.org/abs/2401.12467 (2024).

[6] Ge, Y. Q., Yang, Z. X., Zhang, Z. et al. OBWX: A dataset for classification of ancient Chinese ideographs from four historical periods. Zenodo https://doi.org/10.5281/zenodo.20037584 (2026).

[7] Guan, H. S., Yang, H. X., Wang, X. Y. et al. Deciphering oracle bone language with diffusion models. In *Proc. 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*; 10.18653/v1/2024.acl-long.831 (2024).

[8] Wang, P. J., Zhang, K. L., Wang, X. Y. et al. Puzzle pieces picker: Deciphering ancient Chinese characters with radical reconstruction. In *Proc. 17th International Conference on Document Analysis and Recognition (ICDAR)*; 10.1007/978-3-031-70533-5_11 (2024).

[9] Hua, W. J., Nguyen, H. H. & Ge, G. Y. BIRD: Bronze inscription restoration and dating. In *Proc. 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; 10.18653/v1/2025.emnlp-main.1073 (2025).

[10] Sun, J. W., Guan, S. Y., Chen, X. Y. et al. Adaptive difficulty-aware decoupled learning for long-tailed oracle bone character recognition (in Chinese). *J. Electron. Inf. Technol.*; 10.11999/JEIT260327 (2026).

[11] Chen, Z. J., Chen, T. Z., Zhang, W. J. et al. OBI-Bench: Can LMMs aid in study of ancient script on oracle bones?. In *Proc. 13th International Conference on Learning Representations (ICLR)*; 10.48550/arXiv.2412.01175 (2025).

[12] Qiao, R. Q., Tan, Q. N., Dong, G. T. et al. V-Oracle: Making progressive reasoning in deciphering oracle bones for you

and me. In *Proc. 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*; 10.18653/v1/2025.acl-long.986 (2025).

[13] Wu, R., Wang, S. B., Li, X. S. et al. A text image dual conditional stable diffusion model for oracle bone inscription decipherment. *npj Herit. Sci.* **13,** 193; 10.1038/s40494-025-02019-9 (2025).

[14] Li, C., Ding, Z. M., Hu, X. et al. OracleFusion: Assisting the decipherment of oracle bone script with structurally constrained semantic typography. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)* 19893–19902; 10.1109/iccv51701.2025.01850 (2025).

[15] Li, J., Chi, X. K., Wang, Q. F. et al. A comprehensive survey of oracle character recognition: challenges, datasets, methodology, and beyond. *Pattern Recognit.* **169,** 111824; 10.1016/j.patcog.2025.111824 (2026).

[16] Chen, Z. J., Hua, W. J., Li, J. H. et al. Oracle bone inscriptions information processing: a comprehensive survey. *npj Herit. Sci.* **14,** 220; 10.1038/s40494-026-02511-w (2026).

[17] Guo, J., Wang, C., Roman-Rangel, E. et al. Building hierarchical representations for oracle character and sketch recognition. *IEEE Trans. Image Process.* **25,** 104–118; 10.1109/TIP.2015.2500019 (2016).

[18] Huang, S. P., Wang, H. B., Liu, Y. G. et al. OBC306: a large-scale oracle bone character recognition dataset. In *Proc. 15th International Conference on Document Analysis and Recognition (ICDAR)* 681–688; 10.1109/icdar.2019.00114 (2019).

[19] Zhang, Y. K., Zhang, H., Liu, Y. G. et al. Oracle character recognition by nearest neighbor classification with deep metric learning. In *Proc. 15th International Conference on Document Analysis and Recognition (ICDAR)* 309–314; 10.1109/icdar.2019.00057 (2019).

[20] Li, B., Dai, Q. W., Gao, F. et al. HWOBC: a handwriting oracle bone character recognition database. *J. Phys.: Conf. Ser.* **1651,** 012050; 10.1088/1742-6596/1651/1/012050 (2020).

[21] Wang, M. & Deng, W. H. A dataset of oracle characters for benchmarking machine learning algorithms. *Sci. Data* **11,** 87; 10.1038/s41597-024-02933-w (2024).

[22] Jiao, Q. J., Wu, J. W., Liu, Q. et al. A graph-based evolutionary dataset for oracle bone characters from inscriptions to modern Chinese scripts. *npj Herit. Sci.* **13,** 369; 10.1038/s40494-025-01951-0 (2025).

[23] Gao, M. & Tu, B. K. *Gu Wen Zi Lei Bian (Revised Edition)* (Shanghai Classics Publishing House, 2008). (in Chinese).

[24] Li, S. K. *Chu Wen Zi Bian* (East China Normal University Press, 2003). (in Chinese).

[25] Tang, Y. H. *Zhan Guo Wen Zi Bian* (Fujian People's Publishing House, 2001). (in Chinese).

[26] Dong, L. C. *Xin Jin Wen Bian* (The Writers' Publishing House, 2011). (in Chinese).

[27] Luo, Q. M. et al. Entropy-aware structural alignment for zero-shot handwritten Chinese character recognition. *Pattern Recognit.* **180,** 114156; 10.1016/j.patcog.2026.114156 (2026).

[28] Ke, W. J. et al. Large language models in document intelligence: a comprehensive survey, recent advances, challenges, and future trends. *ACM Trans. Inf. Syst.* **44,** 18; 10.1145/3768156 (2026).

[29] He, K. M., Zhang, X. Y., Ren, S. Q. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778; 10.1109/cvpr.2016.90 (2016).

[30] Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2261–2269; 10.1109/cvpr.2017.243 (2017).

[31] Dosovitskiy, A., Beyer, L., Kolesnikov, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. 9th International Conference on Learning Representations (ICLR)*; 10.48550/arXiv.2010.11929 (2021).

[32] Zhou, R. X., Qiu, P. Q., Zhang, Q., Li, C. T. & Yang, X. Ladder-side mixture of experts adapters for bronze inscription recognition. *npj Herit. Sci.* **14,** 30; 10.1038/s40494-025-02294-6 (2026).

[33] Wang, N. et al. Multi-modal ancient scripts recognition via deep learning with data homogenization and augmentation. *npj Herit. Sci.* **13,** 522; 10.1038/s40494-025-02095-x (2025).

[34] Wang, Y. B., Zuo, Y. N., Jiang, L. X., Qian, C. & Ubul, K. HCT-Net: A hybrid CNN-transformer network for robust oracle bone script recognition. In *Proc. 28th International Conference on Pattern Recognition (ICPR)* 487–502; 10.1007/978-3-032-31654-7_33 (2026).

[35] Zhang, Y. K., Zhang, H., Liu, Y. G. & Liu, C. L. Oracle character recognition based on cross-modal deep metric learning (in Chinese). *Acta Autom. Sinica* **47,** 791–800; 10.16383/j.aas.c200443 (2021).

[36] Wu, L., Liu, G. Y., Liu, P. L., Bin Abdullah, M. I. & Xu, M. L. Detecting unknown and classifying known oracle bone characters via novel data augmentation. *npj Herit. Sci.* **14,** 445; 10.1038/s40494-026-02678-2 (2026).

[37] Ge, Y. L., Zhang, Y. M. & Roh, S.-B. FAIR-Net: A fuzzy autoencoder and interpretable rule-based network for ancient Chinese character recognition. *Sensors* **25,** 5928; 10.3390/s25185928 (2025).

[38] Li, J. W., He, J. R., Wu, J. R. et al. Prism-OBI: A novel framework for oracle bone inscription recognition via visual perception and feature decoupling. *npj Herit. Sci.* **14,** 218; 10.1038/s40494-026-02493-9 (2026).

[39] Hu, Z. P., Cheung, Y.-M., Zhang, Y. et al. Component-level segmentation for oracle bone inscription decipherment. In

Proc. AAAI Conference on Artificial Intelligence 29912–29920; 10.1609/aaai.v39i27.35030 (2025).

[40] Li, J. H., Chen, Z. J., Jiang, R. Z. et al. Mitigating long-tail distribution in oracle bone inscriptions: dataset, model, and benchmark. In *Proc. 33rd ACM International Conference on Multimedia*; 10.1145/3746027.3755067 (2025).

[41] Yang, Z. Y., Han, Z. W., Aysa, A. et al. Long-tailed oracle character recognition based on convolutional neural networks and vision transformers. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 1–5; 10.1109/icassp49660.2025.10888269 (2025).

[42] Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60,** 84–90; 10.1145/3065386 (2017).

[43] Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *Proc. International Conference on Learning Representations (ICLR)*; 10.48550/arXiv.1409.1556 (2015).

[44] Xie, S., Girshick, R., Dollár, P., Tu, Z. W. & He, K. M. Aggregated residual transformations for deep neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1492–1500; 10.1109/cvpr.2017.634 (2017).

[45] Zagoruyko, S. & Komodakis, N. Wide residual networks. In *Proc. British Machine Vision Conference (BMVC)*; 10.5244/c.30.87 (2016).

[46] Sandler, M., Howard, A., Zhu, M. L., Zhmoginov, A. & Chen, L.-C. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 4510–4520; 10.1109/cvpr.2018.00474 (2018).

[47] Howard, A., Sandler, M., Chu, G. et al. Searching for MobileNetV3. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)* 1314–1324; 10.1109/iccv.2019.00140 (2019).

[48] Ma, N. N., Zhang, X. Y., Zheng, H. T. & Sun, J. ShuffleNet V2: Practical guidelines for efficient CNN architecture design. In *Proc. European Conference on Computer Vision (ECCV)* 116–131; 10.1007/978-3-030-01264-9_8 (2018).

[49] Wang, A., Chen, H., Lin, Z. J., Han, J. G. & Ding, G. G. RepViT: Revisiting mobile CNN from ViT perspective. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 15909–15920; 10.1109/cvpr52733.2024.01506 (2024).

[50] Tan, M. X. & Le, Q. V. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proc. 36th International Conference on Machine Learning (ICML)* 6105–6114 (2019).

[51] Tan, M. X. & Le, Q. V. EfficientNetV2: Smaller models and faster training. In *Proc. 38th International Conference on Machine Learning (ICML)* 10096–10106 (2021).

[52] Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K. M. & Dollár, P. Designing network design spaces. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 10428–10437; 10.1109/cvpr42600.2020.01044 (2020).

[53] Liu, Z. et al. A ConvNet for the 2020s. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 11976–11986; 10.1109/cvpr52688.2022.01167 (2022).

[54] Woo, S., Debnath, S., Hu, R. et al. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 16133–16142; 10.1109/cvpr52729.2023.01548 (2023).

[55] Yu, W. H., Zhou, P., Yan, S. C. & Wang, X. C. InceptionNeXt: When Inception meets ConvNeXt. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 10.1109/cvpr52733.2024.00542 (2024).

[56] Dai, Z. H., Liu, H. X., Le, Q. V. & Tan, M. X. CoAtNet: Marrying convolution and attention for all data sizes. In *Advances in Neural Information Processing Systems (NeurIPS)* 3965–3977; 10.48550/arXiv.2106.04803 (2021).

[57] Tu, Z., Xiao, Y., Bagheri, A. et al. MaxViT: Multi-axis vision transformer. In *Proc. European Conference on Computer Vision (ECCV)* 459–479; 10.1007/978-3-031-20053-3_27 (2022).

[58] Mehta, S. & Rastegari, M. Separable self-attention for mobile vision transformers. Preprint at arXiv https://arxiv.org/abs/2206.02680 (2022).

[59] Li, Y. Y., Hu, J., Wen, Y. et al. Rethinking vision transformers for MobileNet size and speed. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)* 16889–16900; 10.1109/iccv51070.2023.01549 (2023).

[60] Liu, Z., Lin, Y. T., Cao, Y. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)* 10012–10022; 10.1109/iccv48922.2021.00986 (2021).

[61] Wang, W. H., Xie, E. Z., Li, X. et al. PVT v2: Improved baselines with pyramid vision transformer. *Comput. Vis. Media* **8,** 415–424; 10.1007/s41095-022-0274-8 (2022).