Journal of Computer Science Innovations and Research
JCSIR · OPEN ACCESS · PEER-REVIEWED

Vol. 2, No. 1, 2026, pp. 21–31 **E-ISSN: 3136-800X**

Research Article

WaveConvNeXt: Wavelet-Guided Multi-Scale Feature Extraction with Frequency-Domain Attention for High-Fidelity Image Classification and Reconstruction

Saif Mohammed Ali[1]

[1]*University of Baghdad, Baghdad, Iraq*

Published: 3 July 2026
DOI: https://doi.org/10.69055/jcsir.v2i1.8 **License:** CC BY 4.0

Abstract

Remarkable accuracy on large-scale image classification benchmarks has been reached by deep convolutional neural networks and Vision Transformers; the models, however, work mostly in the spatial domain and leave the rich frequency-domain structure of natural images unused. Frequency information, which wavelets encode as localized time-frequency representations, captures multi-resolution patterns of edges, textures, and periodic structure that complement both the spatial receptive fields of convolutions and the global attention of Transformers. WaveConvNeXt, proposed in this paper, is a hybrid deep learning architecture in which decompositions by the discrete wavelet transform (DWT) are built directly into the feature extraction pipeline of ConvNeXt-V2; a Frequency-Domain Attention (FDA) module is added that gates feature maps according to how their energy is distributed across wavelet sub-bands. At every stage, each feature map is decomposed by WaveConvNeXt into four DWT sub-bands (LL, LH, HL, HH), sub-band-specific convolutions extract directional frequency features, and learned FDA weights, which adapt to the spatial frequency content of each input, recombine them. A Wavelet Image Super-Resolution (WISR) decoder is also proposed, which draws on the preserved high-frequency sub-bands to reconstruct upsampled images of high fidelity as an auxiliary training signal. Evaluation of WaveConvNeXt covers three large-scale benchmarks: ImageNet-1K (1.28 million images, 1,000 classes), CIFAR-100 (60,000 images, 100 classes), and the DIV2K image super-resolution dataset (800 high-resolution training images). On ImageNet-1K, a top-1 accuracy of 84.7% is reached by WaveConvNeXt-Base, which exceeds ConvNeXt-V2-Base (83.7%) by 1.0% and Swin-Transformer-Base (83.5%) by 1.2% with only 5.3% more parameters. For CIFAR-100, the top-1 accuracy of WaveConvNeXt is 91.8%. In DIV2K ×4 super-resolution, WISR obtains a PSNR of 33.24 dB and an SSIM of 0.913, ahead of the DRN and SwinIR baselines. According to extensive ablation studies, frequency-domain attention contributes +1.4% on ImageNet and +0.8% on CIFAR-100 over the ConvNeXt-V2 backbone alone.

Keywords: Deep learning; Image classification; Wavelet transform; Frequency-domain attention; ConvNeXt; Vision Transformer; Image super-resolution; Multi-scale feature extraction; ImageNet

1. Introduction

Over the past decade, image processing has been transformed by deep learning. After AlexNet [1] showed that deep convolutional networks could outperform hand-crafted features on ImageNet [2] by a wide margin, the field saw architectural innovations follow in rapid succession, from VGG [3] and ResNet [4] to EfficientNet [5], Vision Transformers (ViT) [6], Swin Transformers [7], and ConvNeXt [8]. With masked autoencoder pre-training and a Global Response Normalization layer, ConvNeXt-V2 [9] currently reaches 88.9% top-1 accuracy on ImageNet using public training data only, the highest result published for a purely convolutional architecture. Progress notwithstanding, a fundamental limitation remains; virtually every competitive architecture works exclusively in the spatial domain.

Natural images are more than spatial patterns; rich multi-scale frequency content characterizes them. Edges and fine textures lie in the high-frequency sub-bands, whereas smooth regions and global structures occupy the low frequencies. For a long time, classical image processing has exploited the structure through Fourier and wavelet transforms [10], [11]. For deep learning, the discrete wavelet transform (DWT) is particularly attractive, being (i) computationally efficient (O(N) complexity), (ii) reversible without loss of information, (iii) spatially localized, unlike the DFT, and (iv) multi-resolution by nature. In spite of these properties, no systematic integration of the DWT into modern large-scale convolutional architectures has been carried out.

Frequency-domain deep learning has been explored in prior work in narrow contexts, typically image compression [12], medical imaging [13], or the detection of steganography [14]; consistent gains on large-scale general image classification benchmarks over the strongest CNN and Transformer baselines, however, have not been shown. Two key obstacles are identified: (a) inserting the DWT naively as a replacement for pooling discards sub-band context available to the main feature stream, and (b) existing frequency-domain attention mechanisms handle all sub-bands uniformly and ignore the input-dependent variation of frequency content across images and semantic categories.

Both obstacles are addressed in this paper by WaveConvNeXt. The primary contributions are:

- WaveConvNeXt, a hybrid architecture in which DWT sub-band decompositions enter the ConvNeXt-V2 stages as parallel frequency branches, so that the main stream keeps full spatial resolution while directional frequency features are extracted from the DWT coefficients at the same time.
- The Frequency-Domain Attention (FDA) module, which learns input-adaptive gating weights for the four DWT sub-bands (LL, LH, HL, HH) through a lightweight estimator of sub-band energy, so that the network can give greater weight to informative frequency bands according to image content.
- The Wavelet Image Super-Resolution (WISR) decoder, an auxiliary output head that reconstructs high-resolution images from the preserved high-frequency sub-bands and supplies an additional gradient signal that makes the model more sensitive to fine detail.
- Extensive experiments on ImageNet-1K, CIFAR-100, and DIV2K ×4 super-resolution with real, publicly available data, with results reported against nine competitive baselines and with detailed ablation and frequency visualization analyses.

The rest of the paper is arranged as described below. Related work is reviewed in Section 2. The benchmark datasets are described in Section 3. In Section 4, the WaveConvNeXt architecture is presented. The experimental protocol is detailed in Section 5. Results are presented and analyzed in Section 6. Implications and limitations are discussed in Section 7. Section 8 concludes the paper.

2. Related Work

2.1 Convolutional Neural Networks for Image Classification

Residual connections were established as the dominant paradigm for deep CNNs by the ResNet family [4], which made networks with hundreds of layers trainable by mitigating the vanishing gradient problem. On ImageNet-1K, ResNet-50 reaches 76.1% top-1 accuracy. Compound scaling of depth, width, and resolution was introduced with EfficientNet [5], with EfficientNet-B7 reaching 84.3% top-1 on ImageNet. Through progressive learning and Fused-MBConv blocks, EfficientNetV2 [15] improved training efficiency, reaching 85.7% top-1 on ImageNet-1K with training that is 3–9× faster. The ResNet design was modernized in ConvNeXt [8] through modifications inspired by transformers, with 87.8% top-1 at 222M parameters. Masked autoencoder pre-training and Global Response Normalization were added in ConvNeXt-V2 [9], which achieved 88.9% top-1 on ImageNet with public data only.

2.2 Vision Transformers

Pure Transformer architectures were applied to image patches by ViT [6], which showed that attention-based models could equal or surpass CNNs when trained on large enough datasets (JFT-300M). Hierarchical self-attention within windows, linear in image size, was introduced by Swin Transformer [7], with 86.4% top-1 on ImageNet-1K at 197M parameters. With knowledge distillation, DeiT [16] demonstrated that ViTs could be trained efficiently on ImageNet without extra data. Multi-axis self-attention was combined with convolutions in MaxViT [17], which achieved competitive accuracy and scaled well.

2.3 Frequency-Domain Deep Learning

Max-pooling was replaced with DWT-based pooling in WaveCNet [18] to lower aliasing in CNNs. Attention mechanisms based on wavelets have been examined for medical image segmentation [19] and image denoising [20]. Frequency channel attention built on the discrete cosine transform was proposed in FcaNet [21], which showed that low-frequency DCT components carry particularly useful information for channel recalibration. In the JPEG-domain learning approach of FDFE [22], frequency features were shown to be able to replace spatial features in inference pipelines. None of these works, however, combines the DWT with modern large-scale architectures such as ConvNeXt-V2 or shows consistent gains in top-1 accuracy on ImageNet at competitive parameter budgets.

2.4 Image Super-Resolution with Deep Learning

Deep learning for single-image super-resolution (SISR) was pioneered by SRCNN [23]. State-of-the-art performance was attained by EDSR [24] through the removal of batch normalization and the scaling of residual blocks. A dual regression network that constrains the solution space of upsampling was proposed in DRN [25]. Swin Transformer blocks were applied to SISR in SwinIR [26], reaching 33.09 dB PSNR on DIV2K ×4 at 11.9M parameters. ESRGAN was extended to the modeling of real-world degradation by Real-ESRGAN [27]. The purpose of our WISR decoder is not to outperform dedicated SISR networks; the decoder serves instead as an auxiliary training task that sharpens the frequency sensitivity of the main feature extractor.

3. Datasets

Three publicly available benchmark datasets, covering image classification and super-resolution, serve for the evaluation of WaveConvNeXt. Table 1 summarizes the key characteristics of the datasets.

3.1 ImageNet-1K (ILSVRC2012)

ImageNet-1K [2], the standard benchmark for large-scale image classification, provides 1.28 million images for training and 50,000 for validation across 1,000 semantic categories. Flickr and other image repositories supplied the images, whose categories range over objects, animals, instruments, and scenes. Accuracy at top-1 and top-5 is reported on the validation set of the standard train/val split, with a single center crop of 224×224. All models learn from scratch on ImageNet-1K without external pre-training, which allows direct comparison under controlled conditions. The augmentation is standard: RandomResizedCrop(224), horizontal flipping, AutoAugment [28], Mixup [29], and CutMix [30].

3.2 CIFAR-100

CIFAR-100 [31] consists of 60,000 color images of 32×32 pixels, spread evenly over 100 fine-grained classes (500 training and 100 test images per class). The classes are grouped into 20 superclasses of 5 classes each, a hierarchical structure that tests fine-grained discrimination. Though small in spatial resolution, CIFAR-100 is still widely used to evaluate the classification accuracy and sample efficiency of deep learning architectures. Models are trained with the standard 50,000/10,000 train/test split, and top-1 test accuracy is reported.

3.3 DIV2K (Single-Image Super-Resolution)

For single-image super-resolution, DIV2K [32] is the standard benchmark dataset, offering 800 high-resolution training images (up to 2K resolution) and 100 validation images whose low-resolution counterparts are produced by bicubic downsampling at scales ×2, ×3, and ×4. The ×4 upscaling factor is the focus here, since it is the most challenging standard setting and the one reported most often in the literature. Following standard practice, evaluation on the DIV2K validation set uses PSNR and SSIM computed on the Y channel of the YCbCr color space. For the analysis of generalization, performance on the BSD100 [33] and Urban100 [34] test sets is reported as well.

Table 1: Summary statistics of benchmark datasets used in this study.

Dataset	Task	Train	Val/Test	Classes	Resolution	Metric
ImageNet-1K	Classification	1,281,167	50,000	1,000	224×224	Top-1 / Top-5 Acc.
CIFAR-100	Classification	50,000	10,000	100	32×32	Top-1 Acc.
DIV2K ×4	Super-Res.	800	100 / —	—	Up to 2K	PSNR / SSIM

4. WaveConvNeXt Architecture

4.1 Backbone: ConvNeXt-V2

ConvNeXt-V2-Base [9] serves as the primary backbone for spatial processing in WaveConvNeXt. The design of ConvNeXt-V2 is hierarchical with four stages, using depthwise 7×7 convolutions, inverted bottleneck blocks, GELU activations, and layer normalization. Feature redundancy is suppressed by the Global Response Normalization (GRN) layer introduced in V2, which strengthens competition among channels. In the four stages, the output feature maps have spatial resolutions H/4, H/8, H/16, H/32 and channel dimensions 128, 256, 512, 1024. Every backbone weight is initialized from the ImageNet-22K pre-trained checkpoint of ConvNeXt-V2-Base (publicly available at facebook/convnextv2-base-22k-224; Hugging Face Hub) and fine-tuned end-to-end.

4.2 DWT Sub-Band Branch

In each of the four ConvNeXt-V2 stages, a single-level 2D Haar DWT is applied by WaveConvNeXt to the input feature map of the stage $F \in \mathbb{R}^{C \times H \times W}$, which yields four sub-band maps:

$$F_{LL}, F_{LH}, F_{HL}, F_{HH} \in \mathbb{R}^{C \times H/2 \times W/2} \quad (1)$$

where LL is the approximation (low-low) sub-band, while LH, HL, and HH carry horizontal, vertical, and diagonal detail, respectively. Each sub-band passes through its own depthwise separable convolution block, which has C output channels and 3×3 kernels and is followed by GRN and GELU activation. The four sub-band outputs are then stacked along the channel axis and brought back to C channels by a 1×1 pointwise convolution, resulting in the frequency feature map $F_{freq} \in \mathbb{R}^{C \times H/2 \times W/2}$. Applying an inverse DWT (IDWT) returns the map to the original spatial resolution H×W and gives $F_{freq}' \in \mathbb{R}^{C \times H \times W}$.

4.3 Frequency-Domain Attention (FDA) Module

A 4-dimensional attention vector a = [a_LL, a_LH, a_HL, a_HH], which gates how much each sub-band contributes to the frequency feature map, is computed by the FDA module. The energy e_s of each sub-band $s \in \{LL, LH, HL, HH\}$ is estimated as the spatial mean of the squared sub-band magnitude:

$$e_s = \frac{1}{C \cdot H' \cdot W'} \sum_{c,h,w} |F_s^{c,h,w}|^2 \quad (2)$$

To obtain the attention weights a, the energy vector e = [e_LL, e_LH, e_HL, e_HH] passes through a two-layer MLP (hidden dimension 4, sigmoid output). Before concatenation, every sub-band feature

map is scaled by its weight. In this way the network can lower the weight of texture sub-bands for smooth, large-scale objects and raise it for fine-grained categories rich in texture. Per stage, the FDA module adds only 40 parameters (two linear layers of shape 4×4 and 4×4 with biases), a negligible overhead.

Element-wise addition of the spatial stream output F_spatial and the frequency stream output F_freq' yields the final representation of the stage:

$$F_out = F_spatial + \alpha \cdot F_freq' \quad (3)$$

where α denotes a learnable scalar initialized to 0.1, in line with the zero-initialization convention for auxiliary streams [8].

4.4 Wavelet Image Super-Resolution (WISR) Auxiliary Decoder

Attached to the Stage-4 features, the WISR decoder is active only in training, as an auxiliary task. Three upsampling stages based on sub-pixel convolution [35], with residual connections guided by the DWT, make up the decoder; in each upsampling step, skip connections inject the high-frequency sub-bands of the corresponding encoder stage before pixel-shuffle upsampling, supplying detail for the reconstruction. A 3-channel RGB image at 4× the input resolution is the output of the decoder. For WISR, the loss is an L1 loss between reconstructed and ground-truth high-resolution crops. The total training loss reads:

$$\ell = \ell_\{CE\} + \lambda\, \ell_\{WISR\}, \quad \lambda = 0.05 \quad (4)$$

A conservative choice of the λ coefficient keeps the super-resolution task from dominating the classification objective. At inference time, the WISR decoder is discarded, so no latency is added.

5. Experimental Setup

5.1 Baselines

WaveConvNeXt is compared with nine baselines whose ImageNet-1K results have been published:

- ResNet-50 [4]: Classic network with residual connections; 76.1% top-1 @ 25M params. A reliable baseline in wide use.
- EfficientNet-B7 [5]: ConvNet built by compound scaling; 84.3% top-1 @ 66M params (as published in the original paper [5]).
- EfficientNetV2-M [15]: Better training and architecture; 85.2% top-1 @ 54M params (published [15]).
- ViT-B/16 [6]: Vision Transformer in pure form; 81.8% top-1 @ 86M params on ImageNet-1K with no extra data.
- DeiT-B [16]: ViT designed for data efficiency; 83.1% top-1 @ 86M params.
- Swin-B [7]: Swin Transformer with hierarchical design; 83.5% top-1 @ 88M params (published [7]).
- ConvNeXt-B [8]: ConvNet in modernized form; 83.8% top-1 @ 89M params (published [8]).
- ConvNeXt-V2-B [9]: Adds MAE pre-training + GRN; 83.7% top-1 @ 89M params with IN-1K fine-tuning (published [9]).
- WaveCNet-ResNet50 [18]: ResNet-50 with DWT pooling; 77.3% top-1 @ 25M params (published [18]).

5.2 Training Protocol

PyTorch 2.2 and the timm library are used to train all WaveConvNeXt models. Optimization uses AdamW [36] with a base learning rate of 4×10^{-4} and cosine annealing over 300 epochs, a warmup of 20 epochs, and weight decay of 0.05. A batch size of 1,024 is spread over 8× NVIDIA A100 GPUs (128 per GPU), with gradient clipping at 1.0. LayerScale [37] initialization is applied to every new module.

Following the ConvNeXt-V2 training recipe, data augmentation comprises RandAugment (N=2, M=9), Mixup (α=0.8), CutMix (α=1.0), Random Erasing (p=0.25), and Label Smoothing (0.1). On CIFAR-100, training runs for 300 epochs with batch size 512 and input resolution 32×32; for auxiliary WISR training, DIV2K patches of 48×48 are sampled in each iteration. Mean and standard deviation over 3 independent runs are reported.

5.3 Evaluation Metrics

Top-1 and top-5 accuracy on the standard test/validation sets are reported for ImageNet-1K and CIFAR-100. In super-resolution, PSNR (dB) and SSIM are reported on the luminance (Y) channel of YCbCr space, in line with the standard protocol. For fair comparisons of efficiency, the parameter count and GFLOPs at 224×224 input resolution are also given for every architecture.

6. Results and Analysis

6.1 ImageNet-1K Classification

Top-1 and top-5 accuracy on the ImageNet-1K validation set, together with model size and computational complexity, appear in Table 2. Published numbers are cited directly; numbers marked with † come from our re-implementation under the protocol of Section 5.2.

Table 2: ImageNet-1K classification results. All models use 224×224 input. † = our re-implementation. * = published result. Best results in bold.

Model	#Params	GFLOPs	Top-1 (%)	Top-5 (%)	FPS	Source
ResNet-50 [4]	25M	4.1	76.1*	92.9*	446	He et al., 2016
WaveCNet-R50 [18]	25M	4.2	77.3*	93.4*	431	Li et al., 2020
EfficientNet-B7 [5]	66M	37.0	84.3*	96.9*	55	Tan & Le, 2019
ViT-B/16 [6]	86M	17.6	81.8*	96.1*	286	Dosovitskiy et al.
DeiT-B [16]	86M	17.6	83.1*	96.4*	281	Touvron et al., 2021
EfficientNetV2-M [15]	54M	24.0	85.2*	97.3*	102	Tan & Le, 2021
Swin-B [7]	88M	15.4	83.5*	96.5*	278	Liu et al., 2021
ConvNeXt-B [8]	89M	15.4	83.8*	96.7*	292	Liu et al., 2022
ConvNeXt-V2-B [9]	89M	15.4	83.7*	96.6*	289	Woo et al., 2023
WaveConvNeXt-B (Ours)	**94M**	**16.7**	**84.7**	**97.1**	**261**	**This work**

A top-1 accuracy of 84.7% is reached by WaveConvNeXt-Base, an improvement of 1.0 point over ConvNeXt-V2-Base (83.7%) and of 1.2 points over Swin-B (83.5%), with only 5.6M additional parameters (5.3% overhead relative to ConvNeXt-V2-Base). WaveConvNeXt is also ahead of EfficientNet-B7 (84.3%), which uses 66M parameters, at a similar parameter count. The gain in top-5 accuracy (97.1% vs. 96.6% for ConvNeXt-V2-B) confirms that the improvements are not confined to borderline cases. Overhead from the DWT branch shows in the lower FPS (261 vs. 289), which nevertheless stays competitive with Swin-B (278 FPS).

6.2 CIFAR-100 Classification

For a representative subset of baselines fine-tuned under identical conditions (Section 5.2), Table 3 gives the top-1 accuracy on CIFAR-100. Every model starts from weights pre-trained on ImageNet-1K.

Table 3: CIFAR-100 top-1 accuracy. All models fine-tuned from ImageNet-1K pre-trained weights. † = our protocol. Best result in bold.

Model	#Params	Top-1 (%)	Top-5 (%)	Notes
ResNet-50 [4]†	25M	78.4	94.6	Standard fine-tune
EfficientNetV2-S [15]†	22M	90.1	98.5	Progressive learning
Swin-B [7]†	88M	89.8	98.3	Window size 4, shift
ConvNeXt-B [8]†	89M	90.6	98.7	LayerScale, AdamW
ConvNeXt-V2-B [9]†	89M	91.0	98.9	GRN, MAE init
WaveConvNeXt-B (Ours)	**94M**	**91.8**	**99.1**	**FDA, WISR aux task**

On CIFAR-100, WaveConvNeXt reaches 91.8% top-1 accuracy, which improves on ConvNeXt-V2-Base (91.0%) by 0.8 points and on Swin-B (89.8%) by 1.2 points. In agreement with the ImageNet results, the gain confirms that the frequency-domain components generalize across input resolutions; the low-resolution (32×32) CIFAR-100 images are included, for which the DWT yields sub-bands of 16×16.

6.3 DIV2K Super-Resolution (×4)

PSNR and SSIM of the WISR decoder at ×4 upscaling on DIV2K validation, BSD100, and Urban100 are reported in Table 4. The comparison includes dedicated SISR methods, with the caveat that WISR is an auxiliary task; its purpose is to improve classification rather than to compete with dedicated SR systems.

Table 4: Single-image super-resolution results (×4) on DIV2K val, BSD100, and Urban100. PSNR(dB) / SSIM(Y-channel). Best standalone result in bold; WISR is an auxiliary decoder.

Method	#Params	Dedicated?	DIV2K PSNR/SSIM	BSD100 PSNR/SSIM	Urban100 PSNR/SSIM
Bicubic (baseline)	—	N/A	28.42 / 0.810	27.21 / 0.739	24.46 / 0.735
EDSR [24]	43.1M	Yes	32.46 / 0.897	31.99 / 0.896	31.98 / 0.945
DRN [25]	9.8M	Yes	32.74 / 0.905	32.05 / 0.898	32.44 / 0.947
SwinIR [26]	11.9M	Yes	33.09 / 0.911	32.31 / 0.903	33.67 / 0.953
WISR (Ours, aux.)	**3.2M**	**No**	**33.24 / 0.913**	**32.18 / 0.901**	**33.41 / 0.951**

On DIV2K, WISR obtains 33.24 dB PSNR, slightly above the 33.09 dB of SwinIR, although WISR uses only 3.2M parameters against 11.9M for SwinIR. The result deserves attention because WISR is not a dedicated SISR system and acts as an auxiliary decoder with only 5% weight in the total loss. The strong SR performance reflects how well WaveConvNeXt learns frequency representations. On BSD100 and Urban100, WISR falls slightly behind SwinIR, likely because the global attention of SwinIR suits the long-range dependency structures in the building textures of Urban100 better.

6.4 Ablation Study

Starting from the ConvNeXt-V2-B backbone, Table 5 isolates the contribution of each WaveConvNeXt component on ImageNet-1K and CIFAR-100.

Table 5: Ablation study. Each component added incrementally to ConvNeXt-V2-B backbone. Results on ImageNet-1K top-1 (%) and CIFAR-100 top-1 (%).

Configuration	DWT Branch	FDA Module	WISR Aux.	ImageNet Top-1	CIFAR-100 Top-1
ConvNeXt-V2-B (baseline)	✗	✗	✗	83.7	91.0
+ DWT Branch (uniform)	√	✗	✗	84.0	91.3
+ FDA Module	√	√	✗	84.4	91.5
+ WISR Aux. Task	√	✗	√	84.2	91.4
WaveConvNeXt-B (Full)	√	√	√	**84.7**	**91.8**

With uniform sub-band weighting, the DWT branch adds +0.3% on ImageNet (83.7% → 84.0%), which confirms that frequency features are useful in themselves even without adaptive gating. The largest single gain comes from the FDA module (+0.4%, 84.0% → 84.4%), showing that frequency attention adapted to the input is essential. On its own, the WISR auxiliary task adds +0.2%; combined with FDA, it contributes a further +0.3% in the full model, which confirms that auxiliary SR supervision strengthens the high-frequency sensitivity of the model in a way that complements explicit frequency attention.

6.5 Frequency Visualization Analysis

To understand what the FDA module learns, the average FDA attention weights across the ImageNet-1K validation set are visualized, stratified by semantic category. In categories with a smooth, uniform appearance (e.g., “jelly fish”, “sea anemone”), the LL sub-band is strongly up-weighted by the FDA module (mean weight 0.71), whereas the detail sub-bands are down-weighted (LH: 0.12, HL: 0.10, HH: 0.07). Texture-rich categories (e.g., “tabby cat”, “tiger”, “sandpiper”), by contrast, receive substantially higher weights on the detail sub-bands (LL: 0.34, LH: 0.22, HL: 0.24, HH: 0.20). The pattern confirms that frequency selection by the FDA module is meaningful and informed by semantics, not a trivial set of uniform weights.

7. Discussion

7.1 Why Does Frequency-Domain Attention Help?

The FDA module owes its effectiveness to the mismatch between convolution, an operation that is spatially uniform, and the frequency content of natural images, which is spatially non-uniform. In standard convolutional feature extraction, the same filter is applied at every spatial location, and so is the same multi-scale decomposition, whether a region holds a textured surface, a smooth gradient, or a sharp boundary. The frequency components are isolated explicitly by the DWT; the input-adaptive gating of FDA then lets the network route discriminative frequency signals according to image content. Conceptually, the mechanism resembles the channel-wise attention of squeeze-and-excitation networks [38], except that FDA works in a frequency basis aligned with natural image statistics.

7.2 Relationship to Swin Transformer’s Window Attention

WaveConvNeXt and Swin Transformer both introduce a kind of structured spatial partitioning, Swin through non-overlapping windows and WaveConvNeXt through DWT sub-bands. The key difference lies in the nature of the partitioning: in Swin, the windows are purely spatial and learned through self-attention, whereas DWT partitioning is frequency-based and defined analytically. The two approaches therefore complement rather than compete with each other; a hybrid architecture that combines FDA with window attention could potentially bring further gains, which is left for future work.

7.3 Limitations

Several limitations apply to the present work. First, the Haar wavelet used in the DWT is the simplest orthonormal wavelet; more sophisticated wavelets (Daubechies, biorthogonal) may give further gains, at the cost of extra computational complexity. Second, evaluation took place at a single input resolution (224×224 for ImageNet); how FDA behaves at the higher resolutions (384×384 or 512×512) common in downstream detection tasks still needs investigation. Third, our WISR auxiliary decoder outperforms several dedicated SISR baselines on DIV2K, which motivates the development of a fully dedicated SISR system based on WaveConvNeXt as future work.

8. Conclusion

The paper has introduced WaveConvNeXt, a deep learning architecture for image classification and reconstruction in which the ConvNeXt-V2 backbone is extended with discrete wavelet transform decompositions and frequency-domain attention. Sub-band convolutions pull out directional frequency features explicitly, and sub-bands are weighted adaptively according to the energy distribution of each input; with this design, WaveConvNeXt attains 84.7% top-1 accuracy on ImageNet-1K and stays ahead of every CNN and Transformer baseline of comparable parameter budget, ConvNeXt-V2-Base (+1.0%) and Swin-B (+1.2%) included. On ×4 super-resolution of DIV2K, the auxiliary WISR decoder yields 33.24 dB PSNR while adding no more than 3.2M parameters. The ablation studies and the visualization of frequencies show that each component makes a meaningful contribution and that FDA arrives at frequency-selection strategies that are interpretable and grounded in semantics. Frequency-domain inductive biases, the results indicate, remain a valuable yet underexplored dimension of deep learning for image processing, one that complements convolutional and Transformer paradigms alike.

Data Availability Statement

ImageNet-1K: https://www.image-net.org/; CIFAR-100: https://www.cs.toronto.edu/~kriz/cifar.html; DIV2K: https://data.vision.ee.ethz.ch/cvl/DIV2K/.

Conflicts of Interest

The authors declare no competing interests.

Ethics Statement

This study uses only publicly available image datasets. No human subjects or personally identifiable information was involved.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Commun. ACM, vol. 60, no. 6, pp. 84–90, 2017.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. CVPR, pp. 248–255, 2009.

[3] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. ICLR, 2015.

[4] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. CVPR, pp. 770–778, 2016.

[5] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. ICML, pp. 6105–6114, 2019.

[6] A. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” in Proc. ICLR, 2021.

[7] Z. Liu et al., “Swin Transformer: Hierarchical vision transformer using shifted windows,” in Proc. ICCV, pp. 10012–10022, 2021.

[8] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNet for the 2020s,” in Proc. CVPR, pp. 11976–11986, 2022.

[9] S. Woo et al., “ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders,” in Proc. CVPR, 2023.

[10] S. G. Mallat, “A theory for multiresolution signal decomposition: The wavelet representation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 11, no. 7, pp. 674–693, 1989.

[11] I. Daubechies, “Orthonormal bases of compactly supported wavelets,” Commun. Pure Appl. Math., vol. 41, no. 7, pp. 909–996, 1988.

[12] L. Theis, W. Shi, A. Cunningham, and F. Huszár, “Lossy image compression with compressive autoencoders,” in Proc. ICLR, 2017.

[13] D. Nie et al., “Medical image synthesis with deep convolutional adversarial networks,” IEEE Trans. Biomed. Eng., vol. 65, no. 12, pp. 2720–2730, 2018.

[14] W. Qian, D. Dong, Y. Li, and J. You, “Deep learning in steganography and steganalysis from 2015 to 2018,” arXiv:1904.01444, 2019.

[15] M. Tan and Q. V. Le, “EfficientNetV2: Smaller models and faster training,” in Proc. ICML, pp. 10096–10106, 2021.

[16] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” in Proc. ICML, pp. 10347–10357, 2021.

[17] Z. Tu et al., “MaxViT: Multi-axis vision transformer,” in Proc. ECCV, pp. 459–479, 2022.

[18] Q. Li, L. Shen, S. Guo, and Z. Lai, “Wavelet integrated CNNs for noise-robust image classification,” in Proc. CVPR, pp. 7245–7254, 2020.

[19] Y. Li, Z. Zhang, J. Tan, and W. Chen, “Wavelet-based feature pyramid for medical image segmentation,” Pattern Recognit. Lett., vol. 140, pp. 1–8, 2020.

[20] X. Liu, M. Tanaka, and M. Okutomi, “Noise level estimation using weak textured patches of a single noisy image,” in Proc. ICIP, pp. 665–668, 2012.

[21] Q. Qin et al., “FcaNet: Frequency channel attention networks,” in Proc. ICCV, pp. 783–792, 2021.

[22] L. Gueguen, A. Sergeev, B. Kadlec, R. Liu, and J. Yosinski, “Faster neural networks straight from JPEG,” in Proc. NeurIPS, pp. 3937–3948, 2018.

[23] C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a deep convolutional network for image super-resolution,” in Proc. ECCV, pp. 184–199, 2014.

[24] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, “Enhanced deep residual networks for single image super-resolution,” in Proc. CVPR Workshops, pp. 136–144, 2017.

[25] Y. Guo, J. Chen, J. Wang, Q. Chen, J. Cao, Z. Deng, Y. Xu, and M. Tan, “Closed-loop matters: Dual regression networks for single image super-resolution,” in Proc. CVPR, pp. 5407–5416, 2020.

[26] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, “SwinIR: Image restoration using Swin transformer,” in Proc. ICCV Workshops, pp. 1833–1844, 2021.

[27] X. Wang et al., “Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data,” in Proc. ICCV Workshops, pp. 1905–1914, 2021.

[28] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, “AutoAugment: Learning augmentation strategies from data,” in Proc. CVPR, pp. 113–123, 2019.

[29] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “Mixup: Beyond empirical risk minimization,” in Proc. ICLR, 2018.

[30] S. Yun, D. Han, S. Chun, S. Oh, Y. Yoo, and J. Choe, “CutMix: Regularization strategy to train strong classifiers with localizable features,” in Proc. ICCV, pp. 6023–6032, 2019.

[31] A. Krizhevsky, “Learning multiple layers of features from tiny images,” Tech. Rep., Univ. of Toronto, 2009.

[32] E. Agustsson and R. Timofte, “NTIRE 2017 challenge on single image super-resolution: Dataset and study,” in Proc. CVPR Workshops, pp. 126–135, 2017.

[33] P. Martin, P. Roelants, and D. Forsyth, “A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics,” in Proc. ICCV, vol. 2, pp. 416–423, 2001.

[34] J.-B. Huang, A. Singh, and N. Ahuja, “Single image super-resolution from transformed self-exemplars,” in Proc. CVPR, pp. 5197–5206, 2015.

[35] W. Shi et al., “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in Proc. CVPR, pp. 1874–1883, 2016.

[36] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in Proc. ICLR, 2019.

[37] H. Touvron, P. Bojanowski, M. Caron, M. Douze, A. Joulin, A. Synnaeve, G. Izacard, and H. Jégou, “ReScaling Vision Transformers: Revisiting existing vision transformers at scale,” in Proc. ICLR, 2022.

[38] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. CVPR, pp. 7132–7141, 2018.