Adaptive Wavelet Convolution for Vision Tasks: Coordinate Gating, Strip Frequency Gating, and Pixel-wise Fusion

[Author Name(s) Placeholder]

Abstract—Convolutional neural networks are widely used in object detection, semantic segmentation, and image restoration. However, purely relying on fixed-scale local convolutions often struggles to preserve fine textures while maintaining global structures. Wavelet transforms provide a natural multi-scale and interpretable band decomposition, explicitly separating low-frequency structures from high-frequency details, which facilitates frequency-domain modeling. Existing wavelet convolutions typically adopt relatively static depthwise processing or simple re-scaling on subbands, lacking adaptive control over spatial positions, directional frequency responses, and cross-branch interactions. As a result, the contribution of each frequency band is difficult to adjust dynamically across samples and regions.

We propose an interface-compatible adaptive wavelet convolution module, termed AWTConv2d. While keeping the wavelet analysis/synthesis scaffold unchanged, AWTConv2d introduces three key mechanisms. First, it performs per-channel learnable subband mixing via grouped 1×1 transforms, enabling adaptive reorganization among $\{LL, LH, HL, HH\}$. Second, it integrates coordinate gating and strip frequency gating to modulate subband features conditioned on spatial locations and directional low/high-frequency components. Third, it replaces simple summation with pixel-wise adaptive fusion to softly select between the spatial convolution branch and the wavelet reconstruction branch, improving complementary information aggregation.

Experiments on [Dataset Placeholder: e.g., COCO / DOTA / VisDrone / custom dataset] show that AWTConv2d consistently improves [Metric Placeholder: mAP / AP50 / mIoU / PSNR, etc.] with minimal changes to the overall network architecture, and ablation studies verify the contribution of each component.

Index Terms—Wavelet transform, wavelet convolution, frequency-domain modeling, coordinate gating, attention mechanism, feature fusion

I. Introduction

Convolutional neural networks (CNNs) have achieved remarkable progress in object detection and dense prediction. For instance, the DETR family advances end-to-end detection through global matching paradigms [1], [2], and real-time detectors continue to be improved in both architecture and training strategies [3]. Nevertheless, classical convolution operators are still centered around local receptive fields; their capability to model multi-scale patterns, directional textures, and frequency-band information is often obtained indirectly via deep stacking. In scenes with rich textures, large scale variations, or heavy background clutter, relying only on spatial-domain local modeling with fixed kernels may lead to detail loss or accumulated structural errors.

[Affiliation and contact Placeholder]

Frequency-domain approaches offer an alternative perspective. Compared with spatial-domain convolutions, frequency representations can more explicitly separate low-frequency structures from high-frequency details, providing better controllability for feature extraction and feature fusion. Recent studies have explored dynamic frequency filtering and frequency-aware fusion. For example, FFT-based dynamic filtering learns frequency-domain weights to adaptively mix responses across samples [4], while frequency-aware feature fusion explicitly generates low-pass and high-pass kernels to enhance cross-scale fusion [5]. However, directly adopting FFT in practical detection/segmentation pipelines often faces challenges such as resolution-dependent frequency weights, complicated boundary handling, and non-trivial implementation overhead.

Wavelet transforms combine frequency interpretability with spatial locality. They decompose features into multi-scale pyramids of low-frequency subbands (structures) and high-frequency subbands (textures), and can map processed subbands back to the spatial domain via invertible reconstruction. Wavelet-based convolution modules (e.g., WTConv) provide explicit band pathways while remaining structurally simple [6]. Yet, in many existing implementations, subband processing is relatively static, typically consisting of depthwise operations and fixed re-scaling. Consequently, the importance of different bands is hard to adjust dynamically across locations; directional low/high-frequency responses are not explicitly modeled; and the fusion between the spatial branch and the wavelet branch is commonly a simple summation, which may cause branch interference in early training.

To address these issues, we propose an adaptive wavelet convolution module, AWTConv2d, which enhances both wavelet-subband processing and cross-branch fusion. First, we introduce per-channel learnable subband mixing so that information exchange among $\{LL, LH, HL, HH\}$ is no longer restricted by fixed subband semantics. Second, inspired by coordinate gating [7], we generate spatial gates from normalized coordinates to modulate wavelet-domain features, explicitly modeling location-dependent band contributions. Meanwhile, we introduce strip frequency gating [8] to reorganize directional low/high-frequency

components and strengthen the modeling of elongated textures and edge structures. Finally, we employ pixel-wise adaptive fusion [9] between the spatial convolution branch and the wavelet reconstruction branch, avoiding conflicts caused by naive summation and enabling per-pixel selection of more reliable information.

Our contributions are summarized as follows. We present a plug-and-play adaptive wavelet convolution module that remains interface-compatible with existing wavelet convolutions. Without altering the rest of the backbone, the proposed coordinate gating, strip frequency gating, and pixel-wise fusion significantly improve adaptivity across samples and spatial regions. Extensive experiments on [Dataset Placeholder] with both main results and ablations validate the effectiveness of the proposed design.

II. Related Work

A. Wavelet Transforms and Wavelet Convolutions

Wavelet transforms decompose signals into subband representations across multiple scales and orientations through a set of analysis filters, combining frequency interpretability with spatial locality. Unlike FFT, wavelet decompositions preserve local support in the spatial domain, making them more suitable as modular operators embedded in convolutional networks. In vision tasks, wavelets have long been used for multi-scale representation and detail enhancement. Recently, wavelet analysis/synthesis has also been integrated into convolutional structures. For example, WTConv performs multi-level wavelet decomposition of features, applies lightweight convolutions on subbands at each level, and reconstructs features via inverse transforms, offering an explicit frequency-band pathway with low modification cost [6]. Nevertheless, many existing wavelet convolutions remain relatively static in both subband processing and cross-branch fusion, limiting their adaptivity to complex scenes.

B. Frequency-domain Modeling and Dynamic Filtering

Frequency-domain modeling commonly enhances structure and texture representation by explicitly reweighting or filtering frequency components. Dynamic filtering learns content-adaptive frequency weights, enabling input-dependent filtering responses [4]. Frequency-aware fusion further focuses on low-/high-frequency complementarity across scales, generating low-pass and high-pass kernels to resample features and improve residual enhancement [5]. These works highlight the importance of frequency adaptivity, but they also reveal practical issues when directly adopting FFT or complex resampling operators in detection and dense prediction pipelines, such as resolution-dependent parameterization and increased engineering complexity.

C. Spatial Gating and Position-conditioned Modulation

Position-conditioned modulation maps external conditions (e.g., coordinates or semantic cues) to channel-wise or pixel-wise scaling factors, thereby enabling spatially varying responses. CoordGate proposes to generate gating weights for spatially varying convolutions from coordinate encodings, improving spatial adaptivity while keeping the computation efficient [7]. Related conditional modulation approaches include FiLM [10] and AdaIN [11], which more generally demonstrate that learnable feature modulation can effectively enrich model expressiveness.

D. Attention Mechanisms and Feature Fusion

Attention mechanisms explicitly model channel, spatial, or pixel-wise weights to emphasize informative content. For fusing two-branch or multi-branch features, pixel-wise attention provides a finer-grained selection capability. CGA-style fusion computes channel attention and spatial attention to guide pixel attention, achieving per-pixel soft fusion between two feature streams [9]. This idea is aligned with our motivation: when the spatial convolution branch and the wavelet reconstruction branch exhibit different reliability across regions, pixel-wise fusion can mitigate the interference caused by naive summation.

In summary, our method introduces dynamic frequency modulation and spatial conditioning into the wavelet analysis/synthesis scaffold, and adopts pixel-wise fusion to strengthen cross-branch complementarity, improving adaptivity while preserving plug-and-play usability.

III. Method

This section first reviews the basic form of wavelet convolution and then details the proposed adaptive wavelet convolution module, AWTConv2d. The module is designed to remain interface-compatible with WTConv2d, so it can be used as a drop-in replacement for depthwise separable convolutions or depthwise convolutions in existing detection/segmentation networks.

A. Preliminaries: 2D Discrete Wavelet Analysis and Synthesis

Given an input feature map $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W}$, the 2D discrete wavelet transform can be viewed as a downsampling convolution with a set of fixed analysis filters. The output consists of one low-frequency subband and three high-frequency subbands. We denote wavelet analysis as

$$\mathbf{U} = \mathcal{W}(\mathbf{X}) \in \mathbb{R}^{B\times C\times 4\times \frac{H}{2}\times \frac{W}{2}}, \tag{1}$$

where $\mathbf{U}_{:,:,0,:,:}$ corresponds to the LL subband and $\mathbf{U}_{:,:,1:4,:,:}$ corresponds to the three high-frequency subbands $LH/HL/HH$. Wavelet synthesis (inverse transform) is denoted as

$$\hat{\mathbf{X}} = \mathcal{W}^{-1}(\mathbf{U}) \in \mathbb{R}^{B\times C\times H\times W}. \tag{2}$$

In implementation, $\mathcal{W}$ and $\mathcal{W}^{-1}$ can be realized by grouped convolution and transposed convolution, respectively. The filters are determined by the chosen wavelet basis (e.g., db1/db2) and are kept fixed.

B. Baseline: Subband Processing and Fusion in WTConv2d

The basic pipeline of WTConv2d performs multi-level wavelet decomposition, applies depthwise convolution on subband features, reconstructs features through inverse transforms level by level, and finally adds the reconstruction to a spatial depthwise convolution branch [6]. Its key advantage is an explicit band pathway with invertible reconstruction. However, the subband processing is usually static (depthwise convolution plus fixed scaling), and the fusion is a simple summation, lacking input-adaptive selection mechanisms.

C. AWTConv2d: Adaptive Subband Token Mixer

AWTConv2d keeps the wavelet pyramid scaffold but upgrades the per-level subband processing into a composite operator consisting of local refinement, learnable subband mixing, spatial/directional modulation, and subband attention. For the wavelet output at level l, denoted as $\mathbf{U}^{(l)}$, we first reshape it into

$$\mathbf{T}^{(l)} \in \mathbb{R}^{B\times 4C\times H_l\times W_l}, \tag{3}$$

where $H_l = H/2^l$ and $W_l = W/2^l$.

a) (1) Local refinement within each subband: We apply depthwise convolution to $\mathbf{T}^{(l)}$ to capture local patterns inside each subband, denoted as $\phi_{dw}(\cdot)$.

b) (2) Per-channel learnable subband mixing: To enable information exchange among $\{LL, LH, HL, HH\}$ within the same channel, we introduce a grouped 1×1 transform $\phi_{mix}(\cdot)$ with groups set to C. This learns a $4 \to 4$ linear mixing inside each channel. Compared with mixing across all $4C$ channels, this design avoids unconstrained cross-channel coupling, improving stability and interpretability.

c) (3) Coordinate gating: position-conditioned band modulation: The dependence on low-/high-frequency information typically varies across spatial locations. For example, boundary regions may rely more on high-frequency details, while large smooth background regions may rely more on low-frequency structures. Inspired by CoordGate [7], we use the normalized coordinate grid $(x, y) \in [-1, 1]^2$ as conditioning input, and employ a lightweight 1×1 convolutional network to produce a pixel-wise gate $\mathbf{G}^{(l)}_{coord} \in (0, 1)^{B\times 4C\times H_l\times W_l}$. Then we modulate the features by

$$ilde\mathbf{T}^{(l)} = \mathbf{T}^{(l)} \odot \mathbf{G}^{(l)}_{coord}. \tag{4}$$

This implementation does not depend on a fixed input resolution and thus can be applied across stages and varying input sizes.

d) (4) Strip frequency gating: directional low/high-frequency re-organization: To explicitly model directional textures and elongated structures, we introduce strip-based low/high decomposition [8]. Concretely, we perform strip average pooling along the horizontal and vertical directions to obtain low-frequency components, and compute high-frequency components as residuals. Learnable coefficients then mix low/high components and are injected back in a residual manner. Denoting this operator as $\psi_{strip}(\cdot)$, we have

$$\bar{\mathbf{T}}^{(l)} = \psi_{strip}(\tilde{\mathbf{T}}^{(l)}). \tag{5}$$

Unlike relying solely on convolution kernels to implicitly learn directional responses, this mechanism provides structured control over low-frequency smoothing and high-frequency edge emphasis.

e) (5) Subband attention: content-adaptive reweighting across subbands: After subband mixing and spatial/directional modulation, we apply subband attention to $\bar{\mathbf{T}}^{(l)}$. The attention extracts global statistics via global average pooling, and uses a grouped 1×1 transform with groups=C to generate four subband weights per channel, enabling content-adaptive selection among $LL/LH/HL/HH$.

Overall, the level-l subband processing can be summarized as

$$\mathbf{T}^{(l)} \leftarrow \alpha\big(\psi_{strip}(\phi_{coord}(\phi_{mix}(\phi_{dw}(\mathbf{T}^{(l)}))))\big), \tag{6}$$

where $\alpha(\cdot)$ denotes the combination of subband attention and scaling.

D. Wavelet Reconstruction and Pixel-wise Cross-branch Fusion

After processing L levels of subbands, we perform top-down inverse transforms to reconstruct the wavelet branch output $\mathbf{X}_{wave}$. In parallel, the spatial branch produces $\mathbf{X}_{base}$ using depthwise convolution.

Prior approaches often fuse branches by a direct sum $\mathbf{X}_{base} + \mathbf{X}_{wave}$, but the reliability of the two branches can vary across regions. Therefore, we introduce a pixel-wise fusion module. Inspired by content-guided attention fusion [9], we jointly use channel attention and spatial attention to generate a pixel gate $\mathbf{P} \in (0, 1)^{B\times C\times H\times W}$, and perform per-pixel soft fusion:

$$\mathbf{X}_{fuse} = \mathbf{X}_{init} + \mathbf{P} \odot \mathbf{X}_{base} + (1-\mathbf{P}) \odot \mathbf{X}_{wave}, \quad \mathbf{X}_{init} = \mathbf{X}_{base} + \mathbf{X}_{wave}. \tag{7}$$

In addition, we keep a conservative channel-wise gate $\sigma(\mathbf{g})$ to modulate the magnitude of $\mathbf{X}_{wave}$ before fusion, improving training stability when the wavelet branch is initially under-optimized. The final output uses a 1×1 projection to match the channel dimension required by the downstream network.

E. Complexity and Plug-and-play Discussion

Compared with WTConv2d, the added overhead of AWTConv2d mainly comes from the grouped 1×1 subband mixing, the lightweight coordinate gating network, strip pooling operations, and the depthwise 7×7 convolution used in pixel-wise fusion. All components can be efficiently implemented with standard deep learning operators without custom CUDA kernels. Meanwhile, AWTConv2d preserves the same input/output interface as WTConv2d, allowing straightforward replacement and ablation in existing codebases.

IV. Experiments

This section describes the experimental setup, compared methods, and result analysis. Since our contribution is primarily a plug-and-play module replacement, we keep the backbone, training schedule, and data processing pipeline unchanged as much as possible, and only replace the designated convolution modules to evaluate the general performance gain of AWTConv2d.

A. Experimental Setup

a) Datasets and metrics: We evaluate on [Dataset Placeholder: e.g., COCO/DOTA/VisDrone/custom dataset]. For object detection, we report [Metric Placeholder: mAP@0.5:0.95, AP50, AP75]; for segmentation, [Metric Placeholder: mIoU]; and for restoration, [Metric Placeholder: PSNR/SSIM]. The dataset split and evaluation protocol follow [Protocol Placeholder: official split/custom split].

b) Networks and replacement strategy: We choose [Backbone/Detector Placeholder: e.g., RT-DETR/Deformable DETR/YOLO family] as the baseline. The replacement strategy is to substitute the original depthwise convolution or WTConv2d with AWTConv2d at [Location Placeholder: e.g., specific depthwise conv layers in the backbone, certain conv layers in the neck], while keeping the rest of the architecture unchanged. For fair comparison, all hyperparameters other than the replaced module remain the same.

c) Training details: We train with [Optimizer Placeholder: SGD/AdamW], an initial learning rate of [lr Placeholder], weight decay [wd Placeholder], batch size [bs Placeholder], and [epoch Placeholder] epochs. Data augmentation includes [Augmentation Placeholder: multi-scale, random crop, mixup, etc.]. Experiments are conducted on [Hardware Placeholder: GPU model and memory].

B. Compared Methods

We compare against the baseline convolution (Depthwise Conv), WTConv2d [6], and several plug-and-play attention/fusion modules (e.g., coordinate gating [7], strip attention [8], and pixel-wise fusion [9]). All comparisons use the same training schedule and inference settings.

C. Main Results

Table I reports the main results on [Dataset Placeholder]. Under controlled parameter and computation changes, AWTConv2d consistently improves over both the baseline and WTConv2d. This indicates that coordinate gating, strip frequency gating, and pixel-wise fusion effectively enhance the adaptivity of the wavelet branch and the complementarity between the two branches.

TABLE I
Main results on [Dataset Placeholder] (numbers are placeholders and should be filled with actual results).

oprule Method	Params (M)	FLOPs (G)	[Metric Placeholder]
Baseline (DWConv)	[#]	[#]	[#]
WTConv2d [6]	[#]	[#]	[#]
AWTConv2d (Ours)	[#]	[#]	[#]

TABLE II
Ablation of AWTConv2d components (numbers are placeholders and should be filled with actual results).

oprule Setting	SubbandMix	CoordGate	StripGa
A0: WT scaffold + subband DWConv			
A1: + subband mixing	✓		
A2: + coordinate gating	✓	✓	
A3: + strip frequency gating	✓	✓	✓
A4: + pixel-wise fusion (final)	✓	✓	✓

D. Ablation Study

To analyze the contribution of each component, we progressively add the key designs of AWTConv2d while keeping all other settings fixed. Table II presents the ablation results. Overall, subband mixing and subband attention provide a robust baseline gain; coordinate gating further improves location-dependent adaptivity; strip frequency gating is particularly helpful for samples with directional textures and edge structures; and pixel-wise fusion significantly improves local complementarity between the spatial branch and the wavelet branch, leading to the best overall performance.

E. Qualitative Analysis and Discussion (Optional)

To better understand the behavior of the module, we suggest visualizing subband attention weights, the spatial distribution of coordinate gates, and the heatmap of the pixel fusion gate $\mathbf{P}$. For detection, one may further group samples by small objects, elongated objects, and complex backgrounds to examine the impact of strip frequency gating on directional textures. Such visualizations can be included in [Visualization Placeholder: figure id / appendix location].

F. Information Needed to Replace Placeholders

To replace placeholders with publishable experimental numbers, please confirm the dataset name and split, the baseline model and configuration path, the exact layers where the module is replaced, training hyperparameters (lr/epochs/batch size), and the computation/parameter counting protocol (e.g., which script/tool is used to report FLOPs/Params).

V. Conclusion

This paper proposes an adaptive wavelet convolution module, AWTConv2d, targeting plug-and-play integration of wavelet-based operators for vision tasks. While preserving the invertible wavelet analysis/synthesis scaffold,

AWTConv2d improves adaptivity from both subband processing and cross-branch fusion. In the wavelet domain, we introduce per-channel learnable subband mixing and subband attention to enable content-adaptive reorganization among $LL/LH/HL/HH$. We further incorporate coordinate gating for position-conditioned modulation, and employ strip frequency gating to explicitly model directional low/high-frequency components. Finally, we adopt pixel-wise adaptive fusion to softly select between the spatial branch and the wavelet reconstruction branch at a fine granularity, mitigating the interference caused by naive summation.

Experimental results on [Dataset Placeholder] demonstrate that AWTConv2d provides consistent performance gains with a small additional computational cost, and ablation studies verify the effectiveness of each component. Future work may proceed in two directions. First, we will explore stronger forms of dynamic frequency filtering, such as low-rank or interpolatable frequency weights that generalize across resolutions. Second, we will extend the proposed adaptive mechanisms to more general multi-branch architectures and multimodal tasks, further evaluating robustness and interpretability in complex real-world scenarios.

References

[1] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in European Conference on Computer Vision, 2020, pp. 213–229.

[2] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in International Conference on Learning Representations, 2021.

[3] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on real-time object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 16 965–16 974.

[4] [DynamicFilter], "Fft-based dynamic token mixer for vision," arXiv preprint arXiv:2303.03932, 2023.

[5] [FreqFusion], "Frequency-aware feature fusion for dense image prediction," arXiv preprint arXiv:2408.12879, 2024.

[6] [WTConv], "Wavelet transform convolution," arXiv preprint arXiv:2407.05848, 2024.

[7] [CoordGate], "Coordgate: Efficiently computing spatially-varying convolutions in convolutional neural networks," arXiv preprint arXiv:2401.04680, 2024.

[8] [FSA], "Dual-domain strip attention for image restoration," Neural Networks, 2024.

[9] [DEA-Net], "Dea-net: Single image dehazing based on detail enhanced convolution and content-guided attention," in [/ TIP/ACM MM/], 2024.

[10] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in Proceedings of the AAAI Conference on Artificial Intelligence, 2018.

[11] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1501–1510.