LWTformer: A Detail-Aware, Learnable Wavelet-Transformer for Ancient Chinese Character Image Restoration

Supplementary Material

1. Overview

This document is the technical appendix of the main paper, which includes additional introductions to theoretical knowledge and more ablation experiments on the components we proposed. At the end of the document, more interesting experimental visualizations are presented. We hope that this technical appendix will enable you to gain a deeper understanding of the technical scheme proposed in our main paper.

2. 2D-DWT: Theory and Implementation

The Two-Dimensional Discrete Wavelet Transform (2D-DWT) is a fundamental tool for multi-resolution analysis of images. For an input image $f(x, y) \in \mathbb{R}^{H \times W}$, it is implemented efficiently as a separable transform by applying the 1D-DWT sequentially along the rows and columns.

2.1. Decomposition Process

The transformation is performed in two distinct steps:

(i) Row-wise Transformation: First, the 1D-DWT is applied to every row of the image. This operation decomposes each row into its low-frequency and high-frequency components, halving the width of the representation. Mathematically, for a row x, this is expressed as:

$$R_{\text{low}}(x, k) = \sum_{y=0}^{W-1} f(x, y) \cdot \frac{1}{\sqrt{2}} \cdot \phi\left(\frac{y - 2k}{2}\right), \quad (1)$$

$$R_{\text{high}}(x, k) = \sum_{y=0}^{W-1} f(x, y) \cdot \frac{1}{\sqrt{2}} \cdot \psi\left(\frac{y - 2k}{2}\right), \quad (2)$$

where $\phi(\cdot)$ is the scaling function (low-pass filter) that extracts approximations, and $\psi(\cdot)$ is the wavelet function (high-pass filter) that extracts details. The term $\frac{y-2k}{2}$ embodies the core of the wavelet transform: the divisor 2 represents dilation (analysis at a coarser scale), and $2k$ represents translation (shifting the basis). The output index k runs over $W/2$ points due to downsampling.

(ii) Column-wise Transformation: The resulting row-transformed components, R_{low} and R_{high}, are then processed column-by-column using the same 1D-DWT. This step halves the height of the representation and produces the final four subbands, each of size $H/2 \times W/2$.

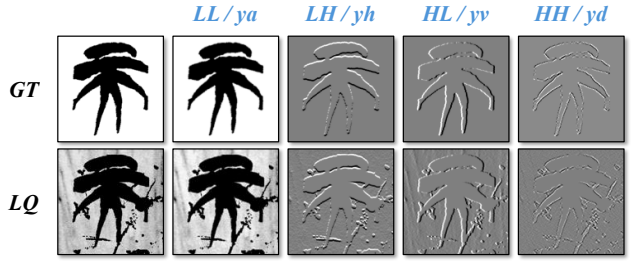

Figure 1. The 2D-DWT results of the ancient character images are presented, with the top and bottom rows corresponding to the high-resolution (*GT*) and low-resolution (*LQ*) versions, respectively.

2.2. Resulting Subbands and Their Interpretation

As shown in Fig. 1, the column-wise transformation on the row components yields four distinct subbands, each capturing specific information:

- LL/ya (Approximation Subband): Obtained by low-pass filtering both rows and columns. It contains the coarse-scale approximation and overall structure of the image.
- LH/yh (Horizontal Detail Subband): Obtained by low-pass filtering rows and high-pass filtering columns. It emphasizes **horizontal edges** and textures.
- HL/yv (Vertical Detail Subband): Obtained by high-pass filtering rows and low-pass filtering columns. It emphasizes **vertical edges** and textures.
- HH/yd (Diagonal Detail Subband): Obtained by high-pass filtering both rows and columns. It captures **diagonal details** and typically contains the finest details, along with most of the image noise.

This multi-resolution, multi-directional decomposition, which cleanly separates structural information from directional details, makes the 2D-DWT exceptionally powerful for tasks like image compression (e.g., JPEG 2000), denoising, and feature extraction.

2.3. Implementation of Learnable Wavelet Filters

A core component of our proposed LWTformer is the integration of a learnable 2D Discrete Wavelet Transform (DWT). Unlike conventional approaches that rely on fixed, predefined wavelet bases (such as Haar or Daubechies), we parameterize the decomposition filters to allow data-driven adaptation.

Filter Initialization and Parameterization. Let $\mathbf{w}_{lo}$ and $\mathbf{w}_{hi}$ denote the low-pass and high-pass decomposition filters, respectively. In our implementation, these filters are

instantiated as PyTorch learnable parameters. To ensure training stability and provide a valid initial frequency separation, we employ a warm-start strategy: the filters are initialized with weights corresponding to a standard wavelet basis (e.g., Haar) rather than random Gaussian initialization. Formally, the learnable DWT operation on an input feature map $\mathbf{X}$ is defined as:

$$\mathcal{W}(\mathbf{X}; \mathbf{w}_{lo}, \mathbf{w}_{hi}) \rightarrow \{\mathbf{X}_{LL}, \mathbf{X}_{LH}, \mathbf{X}_{HL}, \mathbf{X}_{HH}\}, \quad (3)$$

where the filter weights $\mathbf{w}_{lo}$ and $\mathbf{w}_{hi}$ are updated via backpropagation alongside the network parameters to minimize the global restoration loss.

Global Filter Sharing. To enforce consistent multi-resolution analysis and maintain parameter efficiency, we implement a global sharing mechanism. The filter parameters $\mathbf{w}_{lo}$ and $\mathbf{w}_{hi}$ are instantiated once within the top-level model architecture. These shared tensors are subsequently passed by reference to every wavelet-based module in the network, including the *Wavelet-Aware Convolutional Gated Attention* (WACGA) modules within the Transformer blocks and the *WaveletDownsample* layers in the encoder. This design ensures that the network learns a unified frequency decomposition scheme across different network depths, while significantly reducing the parameter overhead compared to layer-specific filter learning.

3. Why V is Chosen as the Gating Signal

In both our proposed Spatial-Enhanced Attention (**SEA**) and Wavelet-Aware Convolutional Gated Attention (**WACGA**), V (Value) is selected as the basis for the gating signal. This design is primarily based on the compatibility between the feature attributes of V in the attention mechanism and the function of the gating mechanism.

As shown in Fig. 2, after depth-wise separable convolution, Q, K, and V already contain most of the structural information, with the outlines of characters clearly visible. Therefore, they are fully qualified to serve as gating signals to activate our feature maps.

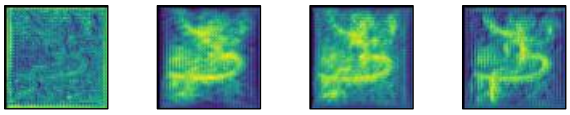

Figure 2. Feature map visualization of Input, Query (Q), Key (K), and Value (V) after depth-wise separable convolution, where character contours are distinct in Q, K, and V.

From the core logic of the attention mechanism, Q (Query) is used to capture query intentions, K (Key) provides the basis for matching, while V directly carries the feature information that needs to be filtered and enhanced. The core goal of the gating mechanism is to dynamically weight effective features and suppress noise, and its target of action is precisely the feature content contained in V. Choosing V as the source of the gating signal enables the gating operation to act directly on the features to be weighted, ensuring the targeted enhancement of key information (such as the slender strokes and damaged edges of ancient characters) and avoiding the dilution of the gating effect caused by introducing irrelevant features (e.g., matching logic information in Q or K).

4. Learnable Directional Operators with Geometric Initialization

In the Wavelet-Aware Convolutional Gated Attention (WACGA) module, we employ a prior-guided initialization strategy to refine the high-frequency wavelet subbands (y_h, y_v, y_d). To explicitly model the directional strokes and fine edge details inherent in ancient characters, we utilize learnable filters initialized with geometric edge detection operators. This integrates a structural inductive bias into the network, providing a geometric "warm-start" that enables the focus on meaningful morphological features from the outset, while retaining the flexibility for data-driven optimization.

4.1. Geometric Warm-Start Priors

We define the initial weights ($\mathbf{W}^{\text{init}}$) for the horizontal, vertical, and diagonal branches using Prewitt-like and Laplacian-based operators. These kernels are formulated as follows:

$$\mathbf{W}_h^{\text{init}} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}, \quad \text{(Horizontal Branch)} \quad (4)$$

$$\mathbf{W}_v^{\text{init}} = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}, \quad \text{(Vertical Branch)} \quad (5)$$

$$\mathbf{W}_d^{\text{init}} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}. \quad \text{(Diagonal Branch)} \quad (6)$$

Horizontal Edge Enhancement (y_h). The horizontal wavelet subband y_h captures intensity variations along the vertical axis, which structurally correspond to horizontal edges or strokes (e.g., the stroke "*Heng*" in Chinese characters). To explicitly enhance these features, we initialize $\mathbf{W}_h$ with the Prewitt-like operator defined above. This operator computes the vertical gradient approximation, effectively highlighting horizontal boundaries. In our architecture, this is preceded by a 1×3 asymmetric convolution to aggregate context along the row dimension without blurring the vertical edges critical for horizontal stroke recognition.

Table 1. Component ablation study across different difficulty levels on the WSC41K dataset. The components are: (a) SEA, (b) WaveDown, (c) WACGA. Best results are in **bold**, second-best are underlined.

Components			Easy				Medium				Hard				Mix			
(a)	(b)	(c)	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
✓	×	×	29.4451	0.9707	0.0926	18.9580	24.9992	0.9397	0.1267	22.4528	19.9636	0.8724	0.1825	31.8432	24.8021	0.9276	0.1339	18.3944
✓	✓	×	29.6894	0.9710	0.0807	16.4397	25.1536	0.9407	0.1136	20.0828	20.0333	0.8745	0.1748	29.5132	24.9582	0.9287	0.1230	16.3413
✓	✓	✓	29.7534	0.9716	0.0763	15.5217	25.2107	0.9417	0.1082	19.7597	20.1095	0.8760	0.1605	28.0930	25.0240	0.9298	0.1150	15.7670
✓	✓	✓	**29.8133**	**0.9722**	**0.0616**	**14.4636**	**25.2964**	**0.9422**	**0.0955**	**18.8708**	**20.1707**	**0.8763**	**0.1530**	**27.6987**	**25.0929**	**0.9302**	**0.1034**	**14.9483**

Vertical Edge Enhancement (y_v). Conversely, the vertical subband y_v contains details regarding vertical edges (e.g., the stroke "*Shu*"). We employ the transposed Prewitt operator $\mathbf{W}_v^{\text{init}}$ to compute the horizontal gradient. This initialization biases the network to detect vertical structures. Accordingly, we pair this with a 3×1 asymmetric convolution in the preceding layer to maintain the integrity of vertical strokes while capturing vertical context.

Diagonal and Point Feature Enhancement (y_d). The diagonal subband y_d typically contains high-frequency variations, including diagonal strokes (e.g., "*Pie*" and "*Na*") and sharp stroke terminals. To capture these omnidirectional high-frequency changes, we initialize $\mathbf{W}_d$ using the Laplacian operator. Unlike the directional Prewitt operators, this second-order derivative kernel is isotropic, making it highly sensitive to rapid intensity changes such as corners and fine details. Additionally, we introduce a Tanh activation in this branch to regulate the magnitude of high-frequency noise often present in the diagonal subband.

Implementation Note. In our implementation, these matrices serve as the initial values for standard convolution layers. Crucially, we enable gradient computation for these layers. This allows the network to fine-tune these geometric priors based on the specific stroke distributions and degradation patterns of ancient Chinese characters, combining the benefits of analytical filter design with data-driven optimization.

5. Extended Ablation Studies

Due to space constraints in the main paper, we present a series of comprehensive extended ablation studies here to thoroughly validate the efficacy and necessity of our proposed architectural components within the LWTformer. These experiments specifically target the contributions of our key innovations, including the Learnable 2D Discrete Wavelet Transform, the Wavelet-Aware Convolutional Gated Attention (WACGA) module, and the efficacy of the geometric prior-guided initialization strategy.

To supplement the main paper's findings, we present an in-depth component ablation study on the WSC41K dataset (Tab. 1). This experiment rigorously evaluates the hierarchical contribution of LWTformer's core modules: (a) SEA (Spatial-Enhanced Attention), (b) WaveDown (Learnable Wavelet Downsample), and (c) WACGA (Wavelet-Aware Convolutional Gated Attention). By sequentially integrating these components and assessing performance across four difficulty levels (Easy, Medium, Hard, and Mix), the results clearly demonstrate the incremental gain, necessity, and synergistic effect of each design in robustly restoring complex ancient character structures and fine details.

Table 2. Ablation study on the weights of frequency-domain loss (λ_1 for $\mathcal{L}_{\text{freq}}$) and perceptual loss ($\lambda_2$ for $\mathcal{L}_{\text{percep}}$) in CharFreqPerceptualLoss on the Oracle Bone Inscription Dataset. Best results are **bolded**, and the second best are underlined.

	$\mathcal{L}_{\text{pix}}$	$\mathcal{L}_{\text{freq}}$	$\mathcal{L}_{\text{percep}}$	PSNR↑	SSIM↑	LPIPS↓	FID↓
(i)	1.0	0.001	0.001	22.9554	0.9528	0.0657	28.6982
(ii)	1.0	0.001	0.005	22.8718	0.9521	0.0641	27.4783
(iii)	1.0	0.001	0.01	22.8776	0.9533	0.0616	25.8578
(iv)	1.0	0.001	0.05	22.7563	0.9519	0.0616	**24.6822**
(v)	1.0	0.005	0.05	22.9827	0.9527	0.0612	26.0083
ours	1.0	0.01	0.05	23.3568	**0.9535**	**0.0610**	26.9066
(vi)	1.0	0.05	0.05	23.5170	0.9479	0.0710	37.1989
(vii)	1.0	0.05	0.1	23.5207	0.9507	0.0652	31.6929
(viii)	1.0	0.1	0.1	**23.6638**	0.9445	0.0737	37.6583

5.1. Ablation Study on CharFreqPerceptualLoss

To determine the optimal balance between pixel-level fidelity and high-level structural and frequency detail, we conduct a sensitivity analysis on the hyper-parameters of our CharFreqPerceptualLoss ($\mathcal{L}_{\text{total}}$). This loss function is a synergistic combination of three constraints—pixel-level ($\mathcal{L}_{\text{pix}}$), frequency-domain ($\mathcal{L}_{\text{freq}}$), and perceptual ($\mathcal{L}_{\text{percep}}$)—and is formally defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pix}} + \lambda_1 \mathcal{L}_{\text{freq}} + \lambda_2 \mathcal{L}_{\text{percep}}, \tag{7}$$

where the pixel loss coefficient is fixed at 1.0. As a complement to the main paper, we first present the component ablation of CharFreqPerceptualLoss on WSC41K (Tab. 3), followed by a detailed sensitivity analysis on the two weighting coefficients (λ_1 and λ_2) shown in Tab. 2 and Tab. 4.

We conduct a detailed ablation study and sensitivity analysis for CharFreqPerceptualLoss across the Oracle

Table 3. Ablation study of components in CharFreqPerceptualLoss on WSC41K. Best results are in **bold**, second-best are underlined.

Loss Config.	Easy				Medium				Hard				Mix			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
only $\mathcal{L}_{\text{pix}}$	29.5407	0.9721	0.0685	**13.7344**	24.9333	**0.9423**	0.1022	24.9557	19.7428	**0.8791**	0.1492	41.2993	24.7385	**0.9311**	0.1066	19.7801
w/o $\mathcal{L}_{\text{freq}}$	29.0455	0.9689	0.0765	16.5566	24.5297	0.9369	0.1084	21.8081	19.4676	0.8699	0.1609	29.2735	24.3471	0.9252	0.1153	17.8407
w/o $\mathcal{L}_{\text{percep}}$	29.7848	**0.9723**	0.0625	15.9580	25.1645	0.9411	0.0963	34.7783	20.0424	0.8764	**0.1468**	59.4100	24.9967	0.9299	**0.1019**	28.1292
all (ours)	**29.8133**	0.9722	**0.0616**	14.4636	**25.2964**	0.9422	**0.0955**	**18.8708**	**20.1707**	0.8763	0.1530	**27.6987**	**25.0929**	0.9302	0.1034	**14.9483**

Table 4. Ablation study on the weight coefficients of frequency-domain loss (λ_1 for $\mathcal{L}_{\text{freq}}$) and perceptual loss ($\lambda_2$ for $\mathcal{L}_{\text{percep}}$) in CharFreqPerceptualLoss on the WSC41K Dataset. Best results are **bolded**.

Case	Loss Weights (λ)			Easy				Medium				Hard				Mix			
	$\mathcal{L}_{\text{pix}}$	$\mathcal{L}_{\text{freq}}$	$\mathcal{L}_{\text{percep}}$	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
(i)	1.0	0.001	0.001	29.5419	0.9713	0.0801	**12.7449**	24.9570	0.9412	0.1147	22.2360	19.7124	**0.8768**	0.1638	39.1068	24.7366	0.9297	0.1195	17.6578
(ii)	1.0	0.001	0.005	29.5449	0.9717	0.0657	15.2371	24.9756	0.9416	0.1078	19.5056	19.7564	0.8764	0.1621	29.1992	24.7584	0.9299	0.1119	17.1725
(iii)	1.0	0.001	0.01	29.6289	0.9716	0.0832	16.5375	25.1113	0.9423	0.1117	21.3791	19.9129	0.8765	0.1582	31.4084	24.8838	0.9301	0.1177	17.8980
(iv)	1.0	0.001	0.05	29.2529	0.9692	0.1002	20.0005	24.7193	0.9379	0.1292	24.9899	19.6552	0.8719	0.1769	31.6986	24.5420	0.9263	0.1354	20.8610
(v)	1.0	0.005	0.05	29.6064	0.9704	0.0979	18.7184	25.0574	0.9400	0.1282	23.5903	19.9458	0.8748	0.1762	32.8342	24.8694	0.9284	0.1341	19.9093
ours	1.0	0.01	0.05	29.8133	0.9722	**0.0616**	14.4636	25.2964	0.9422	**0.0955**	18.8708	20.1707	0.8763	**0.1530**	**27.6987**	25.0929	**0.9302**	**0.1034**	14.9483
(vi)	1.0	0.05	0.05	30.0072	0.9721	0.0630	14.4635	25.4313	0.9420	0.0980	19.4726	20.3565	0.8760	0.1561	28.8692	25.2645	0.9300	0.1057	15.3862
(vii)	1.0	0.05	0.1	29.9804	0.9721	0.0744	14.1512	25.5234	0.9424	0.1111	19.3727	20.4606	0.8760	0.1674	30.0107	25.3210	**0.9302**	0.1176	15.4378
(viii)	1.0	0.1	0.1	**30.1778**	**0.9725**	0.0889	13.5263	**25.6881**	**0.9430**	0.1204	**18.8627**	**20.6038**	0.8752	0.1743	30.8545	**25.4894**	**0.9302**	0.1279	**14.8585**

Bone and WSC41K datasets. First, the component ablation shown in Tab. 3 confirms the crucial synergy of all three loss terms. Removing the perceptual term ($\mathcal{L}_{\text{percep}}$) severely degrades the perceptual quality (high FID), while removing the frequency term ($\mathcal{L}_{\text{freq}}$) compromises overall fidelity. The full loss (*all (ours)*) achieves the best overall balance, notably by securing the lowest LPIPS across all WSC41K difficulty subsets. Second, the sensitivity analysis on the weighting coefficients (λ_1 and λ_2) in Tab. 2 and Tab. 4 reveals that increasing λ_1 mainly boosts PSNR and SSIM, but over-emphasis on frequency loss can deteriorate perceptual quality. Conversely, λ_2 is critical for optimizing LPIPS and FID. Ultimately, our chosen weight configuration, $(\mathbf{1.0}, \mathbf{0.01}, \mathbf{0.05})$, proves to be a robust sweet spot that successfully balances high fidelity with superior perceptual quality across diverse datasets.

5.2. Ablation Study on Adaptivity and Prior Initialization

To validate the efficacy of LWTformer's core design principles, specifically the data-driven adaptivity of the DWT filters and the critical role of structural prior initialization, we conduct a dedicated ablation study. The results, summarized in Tab. 5, systematically compare our proposed mechanisms against conventional or non-adaptive baselines:

- **Filter Adaptivity (Learnable vs. Fixed):** We replace our proposed Learnable Wavelet Filters (LWF) with Fixed Wavelet Filters, utilizing non-learnable coefficients (e.g., fixed Haar basis) for the 2D-DWT operation. This isolates the performance gain attributed solely to the filter's adaptivity.
- **LWF Initialization (Haar Warm-Start vs. Kaiming):** To verify the importance of initializing the LWF with spectral priors, we compare our Haar wavelet-based warm-start strategy against a standard Kaiming uniform initialization, which utilizes random weights.
- **WACGA Initialization (Geometric Warm-Start vs. Standard):** In the WACGA module, we examine the contribution of the geometric edge detection operators used for the 3×3 convolution warm-start (as detailed in Section 4.1). We compare this against a standard 3×3 convolution initialized conventionally (e.g., Kaiming or PyTorch default), thereby assessing the benefit of incorporating explicit spatial structure priors into the attention block.

Table 5. Ablation Study on Adaptivity and Prior Initialization. All configurations are tested on the Oracle Bone Inscription Dataset.

Configuration	PSNR↑	SSIM↑	LPIPS↓	FID↓
w/ Fixed Wavelet Filters	23.2581	0.9524	0.0618	27.1831
w/ Kaiming Init. for LWF	23.1920	0.9505	0.0639	29.0248
w/ Standard Convs in WACGA	23.2407	0.9518	0.0624	28.3233
LWTformer (Ours)	**23.3568**	**0.9535**	**0.0610**	**26.9066**

The results demonstrate that both the learnability of the filters and the introduction of geometric and spectral priors via warm-start initialization are indispensable for achieving LWTformer's superior performance.

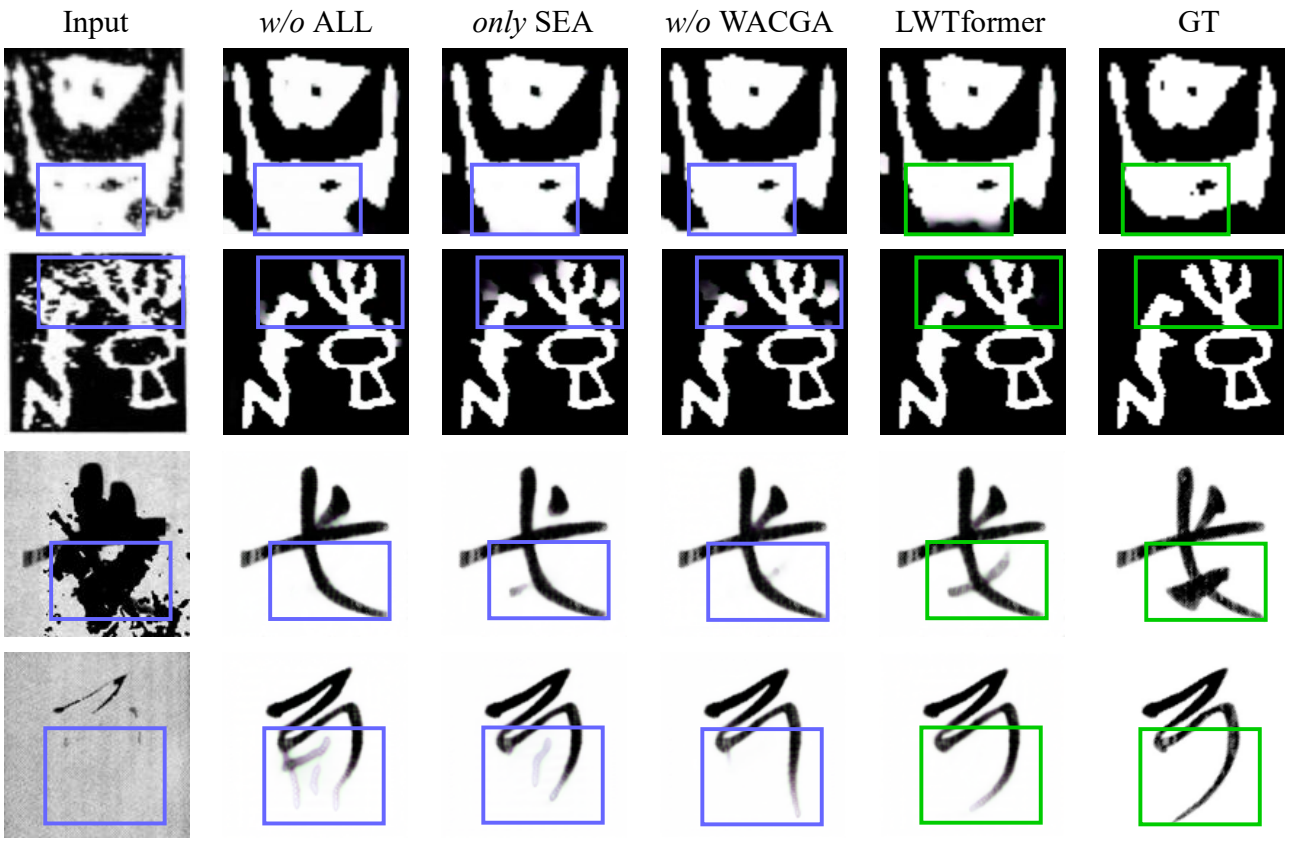

Figure 3. Qualitative Comparison of Ablation Experiments. The figure presents visual results of our ablation study, demonstrating LWTformer's performance under various structural configurations. From left to right, the panels sequentially display: the degraded **Input** image; the baseline configuration removing all proposed components (**w/o ALL**); the configuration utilizing **only SEA** (Spatial-Enhanced Attention); the configuration **w/o WACGA** (Wavelet-Aware Convolutional Gated Attention); the full **LWTformer** model; and the **Ground Truth (GT)**. Low-quality areas are highlighted with purple boxes, and high-quality areas with green boxes.

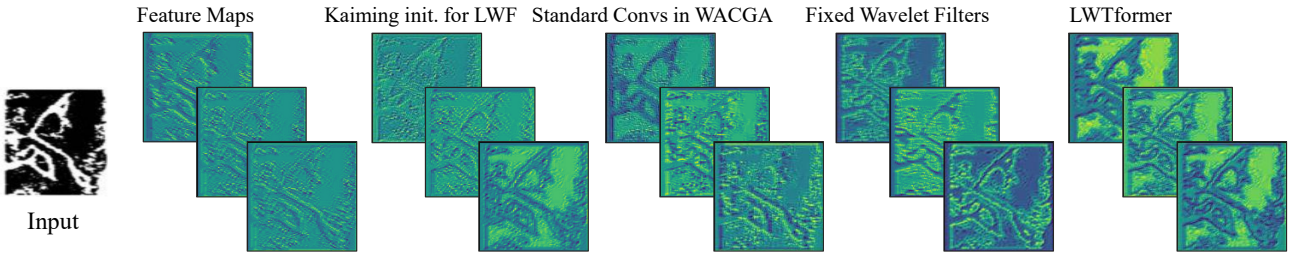

Figure 4. Qualitative Comparison of Key Ablation Components. Panels sequentially display, from left to right: **Input**; an **Intermediate Feature Map** (Full Model); the result of **Kaiming Initialization** (LWT); the result of **Standard Convs in WACGA**; the result of **Fixed Wavelet Filters**; and the full **LWTformer** model.

5.3. Visualization of Ablation Experiments

In Fig. 3, we present **results** from our ablation study on LWTformer's structural configurations. These qualitative comparisons offer confirmation of the contributions made by our core designs. Specifically, we observe that configurations lacking key modules (e.g., the baseline **w/o ALL** and the configuration **w/o WACGA**) suffer from severe detail loss and noise. In contrast, the full **LWTformer** model achieves superior restoration quality, recovering fine strokes and suppressing artifacts. This visual evidence strongly supports the necessity of the **SEA**, **WaveDown**, and **WACGA** modules, aligning perfectly with our quantitative analysis presented in Tab. 1.

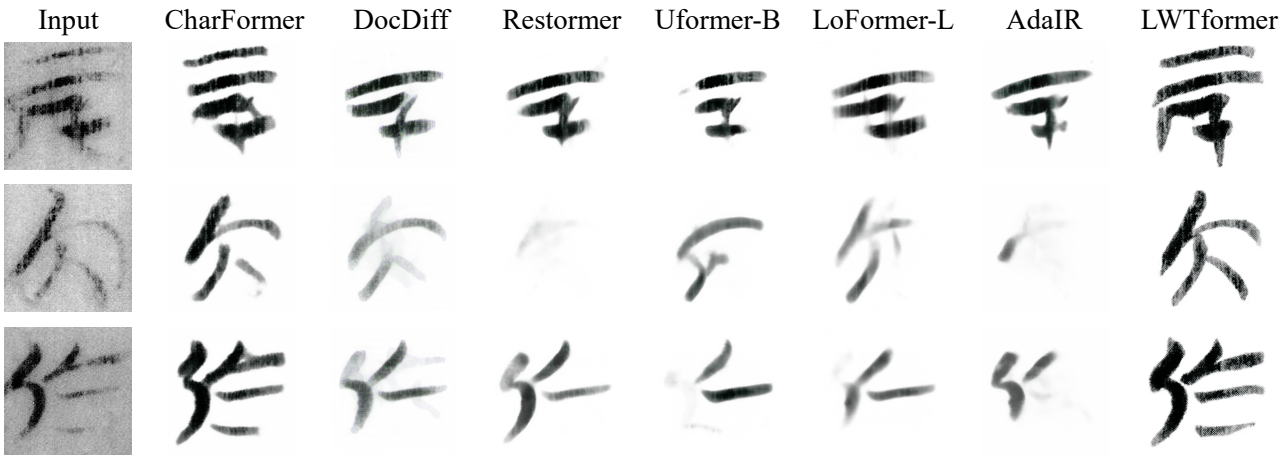

Figure 5. Qualitative Comparison on Real-World Damaged Ancient Character Images. The panels sequentially display, from left to right: the damaged **Input** image; results from several **Competitive Methods**; and the final restoration result achieved by our **LWTformer** model.

Furthermore, intermediate layer visualizations related to the WACGA component ablation are presented in Fig. 4. Compared to the complete LWTformer, all ablated configurations show clear visual flaws. The use of fixed wavelet filters leads to detail blurriness; while the lack of prior-guided initialization (Kaiming Init. and Standard Convs) results in structural artifacts and inaccurate reconstruction. Only the full LWTformer effectively recovers fine strokes with high fidelity. The comparison visually confirms the necessity of learnable filters and prior-guided warm-starts.

6. Real-World Restoration Results

Our model is trained on the synthetic WSC41K dataset and further evaluated on real-world damaged ancient character images collected from Hubei Jian manuscripts. As shown in Fig. 5, existing competitive methods often exhibit blurred strokes, structural distortion, and residual noise when confronted with complex real degradations. In contrast, our LWTformer achieves more visually coherent restorations, effectively recovering fine stroke details and maintaining structural integrity while suppressing background artifacts. These results demonstrate the strong generalization ability and practical applicability of LWTformer for real-world ancient character restoration.

7. Limitations

Although our method demonstrates promising restoration performance, outperforming existing state-of-the-art image restoration approaches, there are still limitations when dealing with severely degraded ancient character images. In particular, our method struggles to produce satisfactory results for severely damaged, rare characters with limited available training examples. These challenging cases often result in incomplete or inaccurate restorations. Additionally, despite the overall high-quality restoration of ancient characters, our LWTformer model still generates some unavoidable artifacts in the restored images. These artifacts, though minimal, can impact the visual quality, particularly in cases with extreme degradation. Furthermore, regarding the proposed CharFreqPerceptualLoss loss function, as discussed in Sec. 5.1, we acknowledge that there is still room for improvement. Refining this loss function may lead to even better performance, particularly in terms of restoring highly degraded characters and reducing residual artifacts.