DocPure: Prompt-Free Unified Document Restoration via Degradation-Aware Structure-Guided Wavelet Modulation

Lingming Su†, Wanglong Lu†, Tao Wang, Kaihao Zhang, Nan Zhang, Liyan An, Hanli Zhao*

***Abstract*—High-quality document images are pivotal for information archiving and downstream automatic processing. However, they are frequently compromised by diverse degradations during uncontrolled acquisition and transmission. While unified document restoration techniques have been proposed to restore images from multiple degradations, they often struggle with training multiple degradation-specific models, reliance on manual task-specific prompts, or cross-task data pairing. To address these limitations, we propose DocPure, a prompt-free unified framework that achieves degradation-aware document restoration. We design a degradation-aware structure autoencoder with degradation-informed routing regularization to predict clean structural priors from degraded inputs. The model is prompt-free at inference, and degradation labels are only used as auxiliary supervision for the routing regularization during training. Furthermore, we introduce a structure-guided wavelet interaction mechanism to bridge frequency-domain features and spatial semantics. Within the structure-guided wavelet interaction mechanism, a cross-frequency adaptive modulation utilizes low-frequency sub-bands to modulate high-frequency recovery, ensuring structural consistency. Extensive experiments demonstrate that DocPure achieves strong performance compared with state-of-the-art methods across various tasks, including deblurring, denoising, compression artifact reduction, and deshadowing. The source code will be made publicly available at https://github.com/LingmingSSS/DocPure.**

***Index Terms*—Image restoration, document image restoration, all-in-one image restoration, prompt-free.**

I. Introduction

Document images frequently suffer from various degradations, such as blur, shadows, and noise. Unlike natural images, document images are characterized by high-contrast strokes and structured layouts [1], making them highly sensitive to such distortions. These degradations not only reduce visual quality, but also severely degrade downstream OCR performance [2], [3]. As an active task of image restoration [4], document image restoration aims to restore degraded document images to their original legibility.

Real-world degradations are diverse and unpredictable [5]–[10]. Given a degraded image, one must typically identify the degradation type first to select the corresponding restoration network. However, existing methods often suffer from notable limitations. As shown in Fig. 1 (a) and Fig. 1 (b), training separate degradation-specific models [11], [12] or relying on manual task-specific prompts [13] is inefficient, while approaches dependent on cross-task data pairing [14] are constrained by data scarcity. Thus, achieving a unified, prompt-free framework capable of handling diverse degradations remains an open challenge.

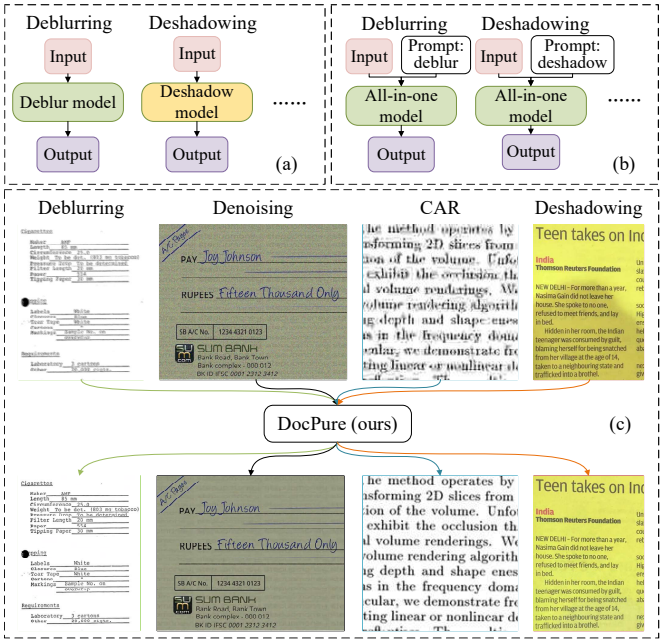

Fig. 1. (a) Single-task document restoration or general-task document restoration models require training multiple models for different degradations. (b) Existing unified document restoration models require manual task-specific prompts about the degradation type at the input. (c) For document images afflicted by diverse degradation types, our method consistently outputs clear, high-fidelity images without requiring manual prompts.

To achieve this, automatically capturing degradation cues and restoring text information effectively are two fundamental problems. As depicted in Fig. 2, we observe that while distinct degradations (*e.g.,* blur, noise, compression, shadow) exhibit complex variations in the spatial domain, they manifest unique and distinguishable spectral signatures in the frequency domain. For instance, blur typically perturbs high-frequency energy, whereas noise introduces anomalous high-frequency responses. Simultaneously, the legibility of document images hinges on high-contrast strokes and sharp layouts [1], which are inherently represented by high-frequency structures.

Lingming Su, Nan Zhang, Liyan An, and Hanli Zhao are with the College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou 325035, China.

Wanglong Lu is with the College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou 325035, China, and also with the AI Analytics Team, Nasdaq, St. John's, NL A1A 0L9, Canada.

Tao Wang is with the vivo Mobile Communication Co., Ltd, Shanghai 201210, China.

Kaihao Zhang is with the College of Engineering and Computer Science, The Australian National University, Canberra, ACT, 2601, Australia.

*Corresponding author: Hanli Zhao (email: hanlizhao@wzu.edu.cn).

†These authors contributed equally to this work.

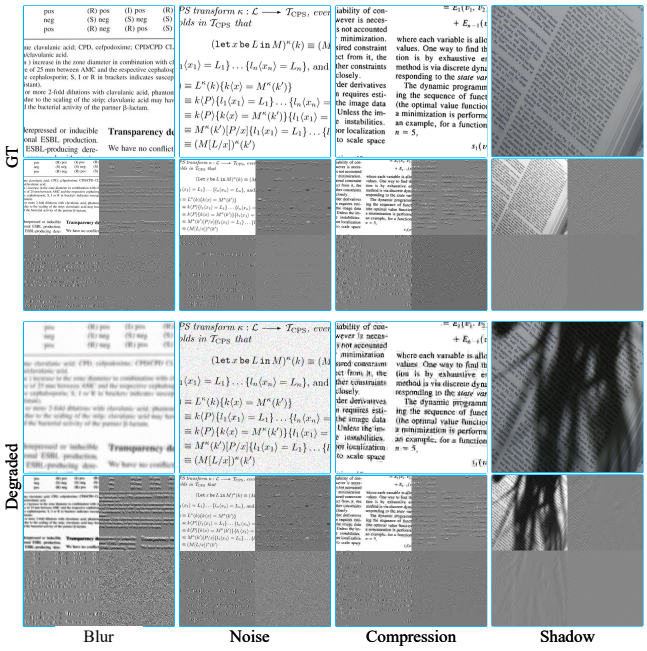

Fig. 2. Visualization of document images and their frequency representations under four typical degradation scenarios: blur, noise, compression, and shadow. The first and third rows display the ground truth (GT) and degraded document images in the spatial domain, while the second and fourth rows illustrate their corresponding frequency representations. Each image is decomposed using the Haar wavelet basis [15], with low-frequency (L) and high-frequency (H) components obtained in each direction, resulting in four feature sub-bands: LL (top-left), HL (top-right), LH (bottom-left), and HH (bottom-right). These degradations exhibit distinct spectral characteristics in the frequency domain, providing informative cues for decoupling of degradation patterns.

To this end, we investigate a two-fold strategy to resolve the above two issues. First, to capture inherent structures without manual prompts, we encode features from degraded images and design a degradation-informed routing regularization to predict clean structural maps. This routing mechanism is trained with degradation-label supervision, enabling the model to learn degradation-discriminative representations during training while operating without prompts at inference. Second, using frequency-domain distributions and predicted structural priors provides internal cues in place of manual prompts, thereby enabling the network to discriminate degradation patterns while preserving semantic integrity.

As shown in Fig. 1 (c), we propose DocPure, a prompt-free unified document restoration framework. We first design a degradation-aware structure auto-encoder (DASAE), a U-Net-style multi-encoder routing network for structure extraction. DASAE employs a degradation-informed routing mechanism to produce input-conditioned routing weights for adaptive structure extraction. This routing mechanism facilitates the prediction of clean structural priors from inputs affected by varying degradations. In addition, by leveraging the distinct spectral signatures of different degradations, we devise a mechanism to facilitate the interaction between spatial semantics and frequency-domain representations, thereby ensuring degradation awareness throughout the restoration process. To realize this, we introduce the structure-guided wavelet interaction mechanism. First, the mechanism decomposes both the degraded input and the predicted structural map into low- and high-frequency sub-bands. Considering that fine-grained details in document images are predominantly manifested in high frequencies, whereas the low-frequency component of the structure map primarily reflects binary regional intensity, we selectively incorporate the high-frequency components to reinforce fine-grained details. Subsequently, we propose the cross-frequency adaptive modulation (CFAM) in the proposed wavelet interaction mechanism to leverage the structural and degradation priors latent in the low-frequency sub-band, which adaptively modulates fine-grained details within the corresponding high-frequency sub-bands. Throughout the process, the mechanism facilitates the interaction between frequency- and spatial-domain features via a cross-attention mechanism, ensuring holistic degradation awareness and structural consistency in the restoration.

Extensive experiments conducted on 12 diverse datasets across four representative degradation scenarios, including document deblurring, denoising, compression artifact reduction, and deshadowing, validate that DocPure achieves performance comparable to or surpassing state-of-the-art (SOTA) approaches, offering a practical solution for document image restoration. In summary, the main contributions of this paper are as follows:

- We introduce DocPure, a novel unified document image restoration framework that achieves adaptive high-quality restoration without requiring manual prompts at inference.
- We propose a degradation-informed routing regularization mechanism to construct a degradation-aware structure auto-encoder that guides feature routing and facilitates structure extraction.
- We introduce a structure-guided wavelet interaction mechanism with cross-frequency adaptive modulation, which leverages spatial-frequency interactions to restore text details while suppressing over-restoration artifacts.

II. Related Work

Recent progress in general image restoration has shown that Transformers are effective in modeling complex degradations, such as non-local artifacts [16], restoration uncertainty [17], and illumination changes [18]. These works suggest the need for degradation-aware modeling, which is also important for document restoration. Existing document restoration research primarily follows two paradigms. Single-task document restoration methods are tailored for specific degradation types, while unified document restoration methods are designed to handle multiple degradations within a general-task or all-in-one architecture.

Single-task document restoration. For **deblurring**, research primarily focuses on recovering high-frequency details under unknown degradation conditions. Blur2sharp [19] generates sharp document images without requiring prior knowledge of the blur kernel. DeblurGAN-CNN [20] improves character legibility by integrating adversarial training with CNN models. DeepDeblur [21] learns both low-level pixel features and high-

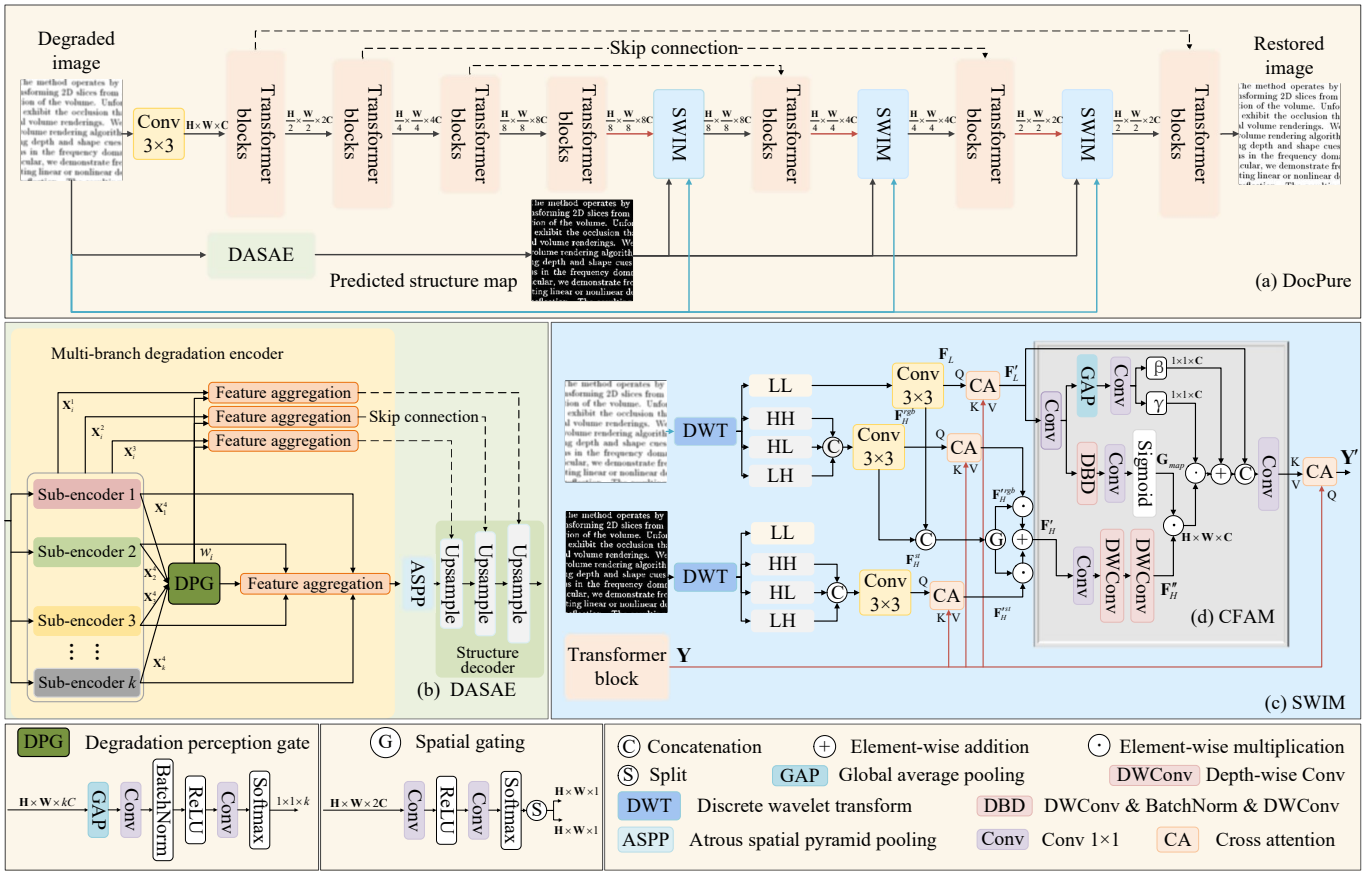

Fig. 3. Overview of the proposed DocPure framework. (a) Our method employs a dual-branch architecture. (b) The degradation-aware structure auto-encoder (DASAE) extracts a clear structure map from the degraded input to guide the Transformer-based restoration network. (c) The structure-guided wavelet interaction module (SWIM) facilitates spatial-frequency interactions to enhance degradation awareness. (d) Cross-frequency adaptive modulation (CFAM) reinforces high-frequency details using low-frequency priors, helping prevent over-restoration artifacts while ensuring structural sharpness.

level semantic features to address complex degradation mappings. In terms of **denoising**, FFDNet [22] handles spatially variant Gaussian noise by taking tunable noise maps as input, and may inadvertently over-smooth fine character strokes. Doc-Attentive-GAN [23] employs an attention-driven GAN to target local reconstruction errors and preserve glyph topology. For **compression artifact reduction**, removing blocking and ringing effects is critical to enhancing the visual quality of text. DMCNN [24] exploits correlations in both pixel and DCT domains to effectively recover information lost during quantization. TPGSR [25] incorporates categorical text priors to guide the restoration process, leveraging semantic guidance to ensure the topological correctness of reconstructed characters. For **deshadowing**, representative methods incorporate physical priors by introducing frequency-aware modules [11], [26] or bilinear imaging models [27]. Recently, attention-oriented detail recovery networks [28] and color-aware strategies [29] have been proposed to enhance texture fidelity after document restoration. To address the bottleneck of the requirement for paired data, studies like FSD [30] explore synthetic data generation schemes based on foreground detection. Although these methods are effective in single-task scenarios, they fail to share knowledge and feature representations across different tasks. In contrast, our prompt-free all-in-one design enables a single prompt-free restoration model to handle multiple degradation types observed during training, without manual prompts or model selection.

General-task document restoration. Initial research focused on designing robust network topologies capable of adapting to various tasks, though often requiring separate training. DE-GAN [31] adopts a paired data-driven supervised learning strategy to achieve the refined removal of both document background shadows and noise. DocDiff [12] introduces diffusion models into document image restoration, achieving robust recovery by predicting the residual between the degraded image and the GT distribution. TextDoctor [32] employs a patch pyramid to achieve low-cost restoration for high-resolution images. Recently, Docstain [33] extracts degradation features at different scales and fuses them using a dual-attention mechanism to enhance restoration performance. However, while effective, these methods typically necessitate training independent parameters for each specific task. Consequently, they lack cross-task feature sharing and joint optimization, failing to achieve true multi-task unification.

All-in-one document restoration. Several all-in-one frameworks have been proposed to handle multiple degradations within a single trained model. DocNLC [14] is a representative outcome of this research. However, its contrastive learning

design necessitates training datasets containing multiple degradations paired with a single GT. This limitation exacerbates the difficulty of data acquisition in real-world applications. DocRes [13] employs a unique prompt generator to produce distinct priors for different degradations, thereby enabling the model to process diverse degradations differentially. Nevertheless, it remains reliant on the manual provision of task-specific prompts to inform the model of the degradation type. UniDocDiff [34] utilizes a unified prior pool to store fixed prior features extracted from input images, aiming to mitigate task interference. However, the model still requires explicit task prompts to select the most appropriate prior from the pool. In contrast, we aim to explore intrinsic features for various degradations to achieve degradation-aware restoration.

III. Method

A. Overview

We present DocPure, a unified framework for document image restoration across diverse degradation scenarios. Unlike prior approaches, DocPure eliminates the need for manual task-specific prompts at inference or cross-task data pairing. Instead, it leverages degradation-informed routing to achieve unified, cross-task restoration within a single model. DocPure directly reconstructs a clean and legible document image from the degraded image.

As illustrated in Fig. 3 (a), the overall architecture of DocPure comprises two core networks: a degradation-aware structure auto-encoder (DASAE) and a restoration network. The backbone of the restoration network adopts a classic 4-level U-shaped encoder-decoder architecture, where each level of the encoder and decoder is composed of multiple Transformer blocks [35] and our structure-guided wavelet interaction modules (SWIM). Given an input degraded document image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ with height H and width W, a 3×3 convolutional layer extracts shallow features from $\mathbf{I}$, which are then forwarded to the backbone. In the meanwhile, DASAE (Fig. 3 (b)) predicts a structure map $\mathbf{I}_M \in \mathbb{R}^{H\times W\times 1}$ from $\mathbf{I}$ by identifying diverse degradation patterns. Then, both the degraded image $\mathbf{I}$ and the predicted structure map $\mathbf{I}_M$ are fed into the restoration network. Within SWIM (Fig. 3 (c)), the interaction between wavelet-domain and spatial-domain features is performed using $\mathbf{I}_M$ and $\mathbf{I}$. It establishes a cross-domain interaction mechanism to explicitly capture degradation characteristics within the spectral domain. As a key module in SWIM, CFAM is designed to balance high-frequency detail enhancement with low-frequency structural constraints. Finally, the network outputs the high-quality restored document image $\hat{\mathbf{I}}$.

B. Degradation-aware structure auto-encoder

Document structure is a critical prior for restoration as it preserves text morphology. However, designing specialized extractors for different degradations is inefficient and limits generalization. To address this, we propose DASAE, which automatically extracts a unified structure prior across diverse degradation types.

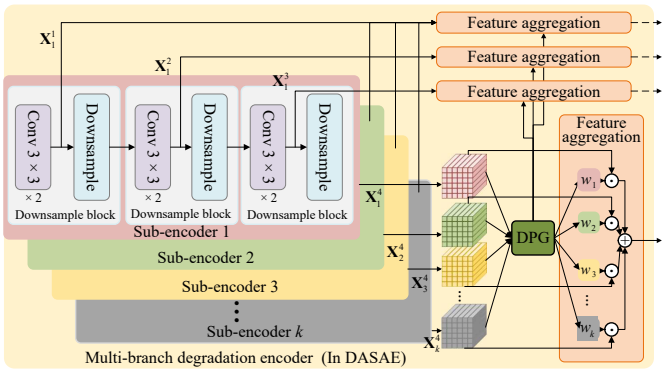

Fig. 4. Details of the multi-branch degradation encoder in DASAE.

As illustrated in Fig. 3 (b), our DASAE first employs a multi-branch degradation encoder with a degradation perception gate (DPG) to model degradation-conditioned features and generate perception confidence scores. These scores then guide feature aggregation, dynamically fusing shallow features into the structure decoder to enhance degradation-aware structure prediction. We further introduce a degradation-informed routing regularization to encourage the routing distribution to be consistent with the degradation-label prior during training.

1) Multi-branch degradation encoder with perception gating: Conventional single encoders often struggle to encompass diverse degradation patterns simultaneously. Conversely, employing multiple encoders to decouple degradation features in parallel offers inherent advantages in achieving task discrimination and feature adaptation within multi-degradation scenarios. Thus, we design a multi-branch encoder gating mechanism, where distinct branches are dedicated to extracting features specific to different degradation types. As shown in Fig. 4, the encoder comprises k sub-encoders, each sub-encoder comprises three downsampling blocks, and each block consists of two convolutional layers and a downsampling layer. The DPG assigns weights to these sub-encoders based on the input features, thereby fusing the multi-path features.

Given a degraded image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, each sub-encoder first extracts its corresponding degradation features. These features are then weighted by the perception confidence score $\mathbf{w} = [w_1, \cdots, w_k]$ generated by the gate. This process can be formulated as:

$$\begin{aligned} &\{\mathbf{X}_i^1, \mathbf{X}_i^2, \mathbf{X}_i^3, \mathbf{X}_i^4\} = \mathrm{SubEncoder}_i(\mathbf{I}), \quad i = 1, \cdots, k, \\ &\mathbf{w} = [w_1, \cdots, w_k] = \mathrm{Softmax}(\mathcal{SE}(\mathbf{X}_1^4, \cdots, \mathbf{X}_k^4)), \end{aligned} \tag{1}$$

where $\mathrm{SubEncoder}_i(\cdot)$ represents the i-th sub-encoder and $\mathbf{w}$ is the perception confidence score. $\mathbf{X}_i^m (m \in \{1, 2, 3\})$ is the feature extracted by the convolutional layer of the m-th downsampling block, and $\mathbf{X}_i^4$ is the feature extracted by the downsampling layer of the third downsampling block. $\mathcal{SE}(\cdot)$ is a Squeeze-and-Excitation block [36]. The perception confidence score $\mathbf{w}$ generated by each sub-encoder will be integrated into the DPG-guided feature aggregation for feature fusion.

To recover spatial details and align features with the degradation pattern, we use a dynamic aggregation strategy guided by a degradation perception gate. In contrast to static skip

connections, our mechanism utilizes the perception confidence score w_i from the DPG as a navigational signal to dynamically synthesize a unified latent representation from the k sub-encoders. Specifically, for both the shallow features at the m-th level and the final deep semantic features of the encoder, we perform a weighted projection:

$$\begin{aligned} \mathbf{X}_{skip}^{m} &= \sum_{i=1}^{k} w_i \cdot \mathbf{X}_i^m, \quad m \in \{1,2,3\}, \\ \mathbf{X}_{task} &= \sum_{i=1}^{k} w_i \cdot \mathbf{X}_i^4, \end{aligned} \tag{2}$$

where $\mathbf{X}_{skip}^{m}$ denotes the aggregated adaptive skip features, which help suppress noise interference irrelevant to the current degradation. $\mathbf{X}_{task}$ represents the deep semantic features fused via the soft routing mechanism of the DPG.

Subsequently, $\mathbf{X}_{task}$ is fed into an Atrous Spatial Pyramid Pooling (ASPP) module [37] to generate $\mathbf{X}_{ASPP}$, further enriching the multi-scale contextual representations of the deep features.

During the subsequent decoding phase, the fused shallow features $\mathbf{X}_{skip}^{m}$ are integrated into the structure decoder. The structure decoder comprises three upsampling blocks. Each block employs transposed convolution to upsample features from the preceding level, thereby restoring spatial resolution. These upsampled features are then concatenated with $\mathbf{X}_{skip}^{m}$ along the channel dimension and subsequently fed into a convolutional module to facilitate feature fusion:

$$\begin{aligned} \mathbf{X}_{dec}^{m} &= \mathcal{F}_{dec}^{m}\left(\left[\mathcal{U}(\mathbf{X}_{in}^{m}), \mathbf{X}_{skip}^{m}\right]\right), \quad m \in \{1,2,3\}, \\ \mathbf{X}_{in}^{m} &= \begin{cases} \mathbf{X}_{dec}^{m+1}, & m \in \{1,2\}; \\ \mathbf{X}_{ASPP}, & m = 3, \end{cases} \end{aligned} \tag{3}$$

where $[\cdot,\cdot]$ denotes the concatenation operation, $\mathcal{U}(\cdot)$ denotes the upsampling operation, and $\mathcal{F}_{dec}^{m}(\cdot)$ represents the feature fusion with two ConvBNReLU operations in the m-th upsampling block. Finally, $\mathbf{X}_{dec}^{1}$ is processed via a 1×1 convolution followed by a sigmoid activation function to generate the predicted structure map $\mathbf{I}_M$.

2) Degradation-informed routing regularization: To prevent the gating block from succumbing to mode collapse [38] during training and encourage degradation-discriminative routing, we introduce a degradation-informed routing regularization.

We formulate external degradation class labels as a distribution of prior knowledge $\mathbf{y} \in \{0,1\}^k$. Our objective is to minimize the Kullback-Leibler (KL) divergence between the generated perception confidence score distribution $\mathbf{w}$ and the target prior distribution $\mathbf{y}$. In this way, degradation labels provide an auxiliary training signal for regularizing the latent routing distribution, encouraging the predicted routing scores to align with the degradation-label prior. Formally, this regularization term is defined as:

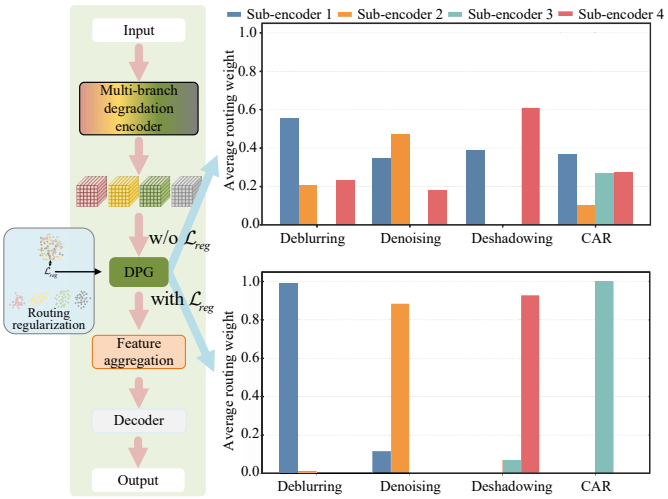

Fig. 5. Visualization of degradation-informed routing regularization. We present the average routing weights across four semantic constraints: deblurring, denoising, CAR, and deshadowing. The visualization demonstrates that our router learns a more sparse and class-selective routing policy, effectively separating distinct degradation patterns. Each sub-encoder is primarily focused on a specific task (*e.g.,* sub-encoder 1 for deblurring, sub-encoder 4 for CAR), verifying the effectiveness of the proposed routing mechanism in distinguishing task semantics.

$$\begin{aligned} \mathcal{L}_{reg} &= \mathcal{D}_{KL}(\mathbf{y}||\mathbf{w}) = \sum_{i=1}^{k} y_i \log \frac{y_i}{w_i} \\ &= \underbrace{\sum_{i=1}^{k} y_i \log y_i}_{\text{Constant}} - \underbrace{\sum_{i=1}^{k} y_i \log w_i}_{\text{Cross-entropy}}. \end{aligned} \tag{4}$$

By imposing this explicit constraint, the DPG is encouraged to learn discriminative degradation features, thereby improving consistency between the routing logic and the input degradation patterns. Since the target distribution $\mathbf{y}$ is deterministic, its entropy $\sum_{i=1}^{k} y_i \log y_i$ is a constant. Consequently, minimizing the KL divergence is mathematically strictly equivalent to minimizing the cross-entropy loss.

We visualize the average routing weight distribution of the DPG within the DASAE module across four tasks in Fig. 5. Without the routing regularization term $\mathcal{L}_{reg}$, the gating block exhibits less separable routing behavior. For specific degradation types, the weight distribution appears relatively dispersed, with multiple encoders being activated simultaneously and lacking significant weight distinctiveness. This failure to separate degradation patterns causes the encoders to learn similar features, which undermines the specialization of individual encoders. With the introduction of $\mathcal{L}_{reg}$, the routing weights exhibit high sparsity and orthogonality. Each degradation task almost exclusively activates a specific dominant encoder. This one-to-one mapping relationship indicates that $\mathcal{L}_{reg}$ successfully compels the DPG to learn discriminative degradation features for different degradation patterns.

3) Loss function of DASAE: To achieve robust performance in terms of both boundary details and overall structure, we formulate a composite loss function. We employ binary cross-

entropy loss $\mathcal{L}_{bce}$ to enhance pixel-level accuracy. To address the inherent class imbalance between the sparse text foreground and the dominant background, we incorporate the Dice loss function ($\mathcal{L}_{dice}$) [39]. As a region-based metric, it has been proven effective in handling segmentation tasks characterized by class imbalance, achieving this by normalizing according to the size of the target region. However, while $\mathcal{L}_{dice}$ ensures region overlap, preserving the fine structural integrity of character strokes necessitates precise boundary localization. To this end, we introduce the boundary loss ($\mathcal{L}_{bd}$), which minimizes the distance between predicted contours and GT contours [40]. By explicitly penalizing boundary deviations, this loss complements the region-based metrics, thereby encouraging the network to generate sharp text structures. The total loss is formulated as follows:

$$\begin{aligned}\mathcal{L} =&\lambda_{bce}\mathcal{L}_{bce} + \lambda_{dice}\mathcal{L}_{dice} + \lambda_{bd}\mathcal{L}_{bd} + \mathcal{L}_{reg}\\ =&\underbrace{-\frac{\lambda_{bce}}{N}\sum_{j=1}^{N}[g_j\log(p_j) + (1-g_j)\log(1-p_j)]}_{\text{BCE loss}}\\ &+\underbrace{\lambda_{dice}\left(1-\frac{2\sum_{j=1}^{N}p_j g_j}{\sum_{j=1}^{N}p_j+\sum_{j=1}^{N}g_j}\right)}_{\text{Dice loss}}\\ &+\underbrace{\frac{\lambda_{bd}}{N}\sum_{j=1}^{N}\phi_{GT(j)}p_j}_{\text{Boundary loss}}+\underbrace{\mathcal{L}_{reg}}_{\text{Routing regularization}},\end{aligned} \tag{5}$$

where N denotes the total number of pixels in the output image, λ_{bce}, λ_{dice}, and λ_{bd} are weights to balance the loss functions. Let $g_j \in \{0,1\}$ and $p_j \in [0,1]$ denote the GT label and the predicted probability for the j-th pixel, respectively. $\phi_{GT(j)}$ denotes the level set function derived from the GT, which corresponds to the signed distance map, *i.e.,* the signed distance from pixel j to the nearest boundary contour.

C. Structure-guided wavelet interaction module (SWIM)

Degradation in document images typically involves loss of high-frequency information and structural discontinuities, imposing stringent demands on the restoration of structures and strokes. Local spatial features alone fail to capture frequency attenuation and global structure, limiting the reconstruction of details. While high-frequency components inherently carry structural information, directly utilizing them from degraded images renders the process susceptible to interference from noise and pseudo-textures. The direct application of predicted structure maps within the spatial domain introduces substantial irrelevant binary redundancy, which can disrupt feature extraction.

As illustrated in Fig. 3 (c), to effectively integrate structural priors with frequency components, we propose a structure-guided wavelet interaction mechanism. This mechanism is implemented as the structure-guided wavelet interaction module (SWIM). Instead of simple feature concatenation, SWIM is conceptually formulated to bridge the semantic gap between spatial layouts and spectral cues. By establishing a collaborative interaction in the wavelet domain, it enables the network to explicitly perceive degradation patterns from spectral signatures while robustly guiding the recovery of fine-grained structural details. Specifically, we employ the Haar wavelet basis [15] to perform wavelet decomposition on the input image $\mathbf{I}$, obtaining a low-frequency component (LL) and three high-frequency components (HL, LH, HH). The low-frequency component is mapped to features $\mathbf{F}_L \in \mathbb{R}^{H\times W\times C}$, while the three high-frequency components are mapped to $\mathbf{F}_H^{rgb} \in \mathbb{R}^{H\times W\times C}$. Meanwhile, the high-frequency components obtained from the wavelet decomposition of the predicted structure map $\mathbf{I}_M$ are mapped to $\mathbf{F}_H^{st} \in \mathbb{R}^{H\times W\times C}$. The process is expressed as:

$$\begin{aligned}&(\mathbf{I}_{LL},\mathbf{I}_{HL},\mathbf{I}_{LH},\mathbf{I}_{HH}) = \mathcal{W}_{Haar}(\mathbf{I}),\\ &\mathbf{F}_L = \mathrm{Conv}_{3\times3}(\mathbf{I}_{LL}),\\ &\mathbf{F}_H^{rgb} = \mathrm{Conv}_{3\times3}([\mathbf{I}_{HL},\mathbf{I}_{LH},\mathbf{I}_{HH}]),\\ &(\mathbf{M}_{LL},\mathbf{M}_{HL},\mathbf{M}_{LH},\mathbf{M}_{HH}) = \mathcal{W}_{Haar}(\mathbf{I}_M),\\ &\mathbf{F}_H^{st} = \mathrm{Conv}_{3\times3}([\mathbf{M}_{HL},\mathbf{M}_{LH},\mathbf{M}_{HH}]),\end{aligned} \tag{6}$$

where $[\cdot,\cdot,\cdot]$ denotes the concatenation operation, $\mathrm{Conv}_{3\times3}(\cdot)$ denotes a 3×3 convolutional layer, and $\mathcal{W}_{Haar}(\cdot)$ signifies the wavelet transform utilizing the Haar wavelet basis for frequency-domain feature extraction.

To inject frequency priors into the semantic space, we employ cross-attention [41], [42]. By utilizing the spatial distribution of frequency features as indices, we retrieve and aggregate relevant semantic contexts from $\mathbf{Y}$, which represents the spatial features extracted by the Transformer block, thereby generating frequency-aware features $\mathbf{F}'_L, \mathbf{F}'^{rgb}_H$, and $\mathbf{F}'^{st}_H$. This operation enables the model to extract multi-scale semantic information aligned with the attentional focus of distinct frequency components.

To balance the influence of image textures and predicted structures, we design a spatial gating module $\mathrm{G}(\cdot)$. Taking the raw frequency features $\mathbf{F}_H^{rgb}$ and $\mathbf{F}_H^{st}$ as input, $\mathrm{G}(\cdot)$ leverages the physical structural information from both low and high frequencies to learn pixel-wise weight maps $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{H\times W\times 1}$. These weight maps are subsequently employed to modulate the frequency-aware features $\mathbf{F}'^{rgb}_H$ and $\mathbf{F}'^{st}_H$ derived in the preceding step:

$$\begin{aligned}&\mathbf{F}'_L = \mathrm{CA}(\mathbf{F}_L,\mathbf{Y},\mathbf{Y}),\\ &\mathbf{F}'^{rgb}_H = \mathrm{CA}(\mathbf{F}_H^{rgb},\mathbf{Y},\mathbf{Y}),\\ &\mathbf{F}'^{st}_H = \mathrm{CA}(\mathbf{F}_H^{st},\mathbf{Y},\mathbf{Y}),\\ &\mathbf{W}_1,\mathbf{W}_2 = \mathrm{G}([\mathbf{F}_H^{rgb},\mathbf{F}_H^{st}]),\\ &\mathbf{F}'_H = \mathrm{Conv}_{1\times1}(\mathbf{W}_1\odot\mathbf{F}'^{rgb}_H + \mathbf{W}_2\odot\mathbf{F}'^{st}_H),\end{aligned} \tag{7}$$

where $\mathrm{CA}(\mathbf{Q},\mathbf{K},\mathbf{V})$ denotes the cross-attention module. $\mathbf{Y}$ serves as the source for both keys ($\mathbf{K}$) and values ($\mathbf{V}$), while the frequency branch features $\mathbf{F}_L$, $\mathbf{F}_H^{rgb}$, and $\mathbf{F}_H^{st}$ independently act as the queries ($\mathbf{Q}$). This design is aimed at preserving the inherent spatial resolution and fine-grained structures of the frequency features. By explicitly injecting the global semantic context from $\mathbf{Y}$ into each frequency component via the attention mechanism [41], it compensates for the semantic deficiency inherent in low-level visual features. The module $\mathrm{G}(\cdot)$ comprises two convolutional layers followed

by ReLU activation, culminating in a Softmax function to generate complementary spatial attention maps. Leveraging the raw frequency structure as a prior, this gating mechanism adaptively adjusts the weights assigned to the high-frequency perceptual components derived from the predicted structure map.

We feed the fused high-frequency perceptual features $\mathbf{F}'_H$ into the CFAM module, where they are adaptively modulated by the low-frequency perceptual features $\mathbf{F}'_L$, producing the enhanced features $\mathbf{F}_{out} \in \mathbb{R}^{H\times W\times C}$:

$$\mathbf{F}_{out} = \mathrm{CFAM}(\mathbf{F}'_H, \mathbf{F}'_L). \tag{8}$$

We then employ the cross-attention module once again to re-integrate the enhanced features $\mathbf{F}_{out}$ into the backbone semantic features $\mathbf{Y}$, thereby updating the feature representation:

$$\mathbf{Y}' = \mathrm{CA}(\mathbf{Y}, \mathbf{F}_{out}, \mathbf{F}_{out}), \tag{9}$$

where we utilize $\mathbf{Y}$ as the query ($\mathbf{Q}$), while $\mathbf{F}_{out}$, which encapsulates rich high-frequency information, serves as both the key ($\mathbf{K}$) and value ($\mathbf{V}$). This design is intended to retrieve and aggregate relevant spatial texture details from the refined frequency-domain features, guided by the backbone semantic features. By propagating the enhanced high-frequency information back into the semantic space, it enriches the detailed representation of features while preserving semantic consistency.

Cross-frequency adaptive modulation (CFAM). Embedded within SWIM, CFAM is designed to further refine feature representations. It is designed to calibrate high-frequency details leveraging the global structural consistency of low-frequency priors.

We propose a low-frequency guided dual modulation mechanism. The high-frequency feature $\mathbf{F}'_H$ undergoes a local enhancement via a 1×1 convolution and two depthwise separable convolutions to obtain $\mathbf{F}''_H$. The low-frequency branch $\mathbf{F}'_L$ branches into two parallel paths to generate modulation parameters. Spatially, it encodes long-range structural contexts to produce an attention map $\mathbf{G}_{map}$, which effectively suppresses noise in high-frequency regions. Meanwhile, in the channel dimension, global statistics of $\mathbf{F}'_L$ are aggregated to generate affine transformation factors $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ for feature recalibration. Finally, $\mathbf{F}''_H$ is spatially gated and channel-wise modulated to achieve restoration:

$$\tilde{\mathbf{F}}_H = \boldsymbol{\gamma} \odot (\mathbf{F}''_H \odot \mathbf{G}_{map}) + \boldsymbol{\beta}, \tag{10}$$

where $\odot$ denotes element-wise multiplication. The calibrated $\tilde{\mathbf{F}}_H$ is then fused with the $\mathbf{F}'_L$ to produce the output.

D. Loss function of restoration network

Following the previous research [35], [42], for the overall restoration network, we adopt the ℓ_1 norm as the training loss function $\mathcal{L}_{\text{rec}}$ to quantify the pixel-wise discrepancy between the restored image $\hat{\mathbf{I}}$ and the GT image $\tilde{\mathbf{I}}$:

$$\mathcal{L}_{\text{rec}} = \|\tilde{\mathbf{I}} - \hat{\mathbf{I}}\|_1. \tag{11}$$

This loss function is simple yet effective, encouraging the model to generate restoration results that are closer to the target image within the pixel space.

TABLE I
SUMMARY OF DATASETS USED IN THIS PAPER.

Source dataset	#Images	Derived dataset	#Images	Usage
Deblurring task				
TDD training set [43]	66,000	TDD training set	40,000	Training
TDD test set [43]	1,600	TDD test set	1,600	Test
FUNSD [44]	199	FUNSD$_{blur}$	199	Test
BCSD [45]	158	BCSD$_{blur}$	158	Test
Denoising task				
TDD training set [43]	66,000	TDD$_{noise}$ training set	40,000	Training
TDD test set [43]	1,600	TDD$_{noise}$ test set	12,800	Test
FUNSD [44]	199	FUNSD$_{noise}$	1,592	Test
BCSD [45]	158	BCSD$_{noise}$	1,264	Test
Compression artifacts reduction task				
TDD training set [43]	66,000	TDD$_{CAR}$ training set	40,000	Training
TDD test set [43]	1,600	TDD$_{CAR}$ test set	17,600	Test
FUNSD [44]	199	FUNSD$_{CAR}$	796	Test
BCSD [45]	158	BCSD$_{CAR}$	632	Test
Deshadowing task				
FSDSRD [30]	14,200	FSDSRD	14,200	Training
RDD training set [29]	4,371	RDD training data	4,371	Training
RDD test set [29]	545	RDD test set	545	Test
OSR [46]	237	OSR	237	Test
Jung et al. [47]	87	Jung et al.	87	Test

IV. EXPERIMENTS

A. Datasets

The datasets used in this paper are summarized in Table I.

Deblurring. For the text deblurring task, we utilized the text deblurring dataset (TDD) [43] as our benchmark. This dataset comprises 66,000 blurry-clean pairs for training and 1,600 samples for testing. Following the convention of recent literature [13], we randomly sampled 40,000 pairs from the TDD training set to train our model and employed its full testing set for evaluation. To assess the generalization capability of the model on unseen document domains, we further tested on two additional datasets with distinct styles: FUNSD [44] and BCSD [45], which contain 199 and 158 GT images, respectively. We synthesized blurry versions using the blur kernels provided by TDD, thereby constructing two extra test sets denoted as FUNSD$_{blur}$ and BCSD$_{blur}$.

Denoising. For the text denoising task, we constructed a noisy document dataset TDD$_{noise}$ based on TDD. For training, we randomly selected 40,000 GT images from the TDD training set and corrupted each image with additive zero-mean Gaussian noise of varying intensities. The standard deviation $\sigma \in \{0, 1, 2, \ldots, 35\}$ was uniformly sampled to cover a wide range of scenarios, from minor interference to severe contamination. The noise was added independently to each RGB channel. For testing, to systematically measure the denoising performance under different noise levels, we constructed hierarchical noise test sets based on the TDD test dataset, FUNSD, and BCSD, denoted as TDD$_{noise}$, FUNSD$_{noise}$, and BCSD$_{noise}$, respectively. Specifically, we added independent and identically distributed (i.i.d.) zero-mean Gaussian noise with $\sigma \in \{0, 5, 10, \ldots, 35\}$ to each GT image.

Compression artifacts reduction (CAR). For the compression artifacts reduction task, we selected 40,000 high-quality GT samples from the TDD training set. By applying JPEG

TABLE II
QUANTITATIVE COMPARISON WITH SOTA DOCUMENT RESTORATION METHODS ON VARIOUS DEGRADATION DATASETS. OUR DOCPURE ACHIEVES THE BEST OVERALL PERFORMANCE. **BOLD** INDICATES THE BEST AND UNDERLINED THE SECOND BEST.

Task	Dataset	Metric	DE-GAN [31] TPAMI'20	DocDiff [12] ACM MM'23	DocStain [33] WACV'25	DocNLC [14] AAAI'24	DocRes [13] CVPR'24	DocPure Ours
Deblurring	TDD	PSNR / SSIM ↑	22.24 / 0.9226	24.00 / 0.9559	30.05 / 0.9861	16.56 / 0.8061	28.34 / 0.9815	**32.55 / 0.9912**
		Char. Acc. / Word Acc. (Paddle) (%) ↑	61.59 / 29.06	77.89 / 44.93	85.94 / 60.34	34.21 / 9.74	83.22 / 55.66	**87.25 / 61.27**
		Char. Acc. / Word Acc. (Nougat) (%) ↑	32.69 / 28.56	37.70 / 29.38	50.45 / 43.05	7.06 / 10.15	47.30 / 37.74	**56.31 / 51.41**
Denoising	TDD_{noise}	PSNR / SSIM ↑	38.32 / 0.9947	37.03 / 0.9912	**43.07 / 0.9987**	19.98 / 0.9047	37.91 / 0.9947	41.26 / 0.9973
		Char. Acc. / Word Acc. (Paddle) (%) ↑	87.62 / 64.85	86.62 / 62.53	**92.41 / 74.15**	63.66 / 19.81	83.89 / 69.13	91.29 / 71.73
		Char. Acc. / Word Acc. (Nougat) (%) ↑	68.16 / 64.93	66.80 / 63.46	**74.95 / 72.39**	4.67 / 12.52	64.84 / 61.37	72.30 / 69.66
CAR	TDD_{CAR}	PSNR / SSIM ↑	18.71 / 0.8871	27.41 / 0.9778	29.37 / 0.9710	18.82 / 0.8837	28.69 / 0.9804	**31.34 / 0.9893**
		Char. Acc. / Word Acc. (Paddle) (%) ↑	70.32 / 28.88	83.89 / 54.56	85.69 / 53.73	55.34 / 16.53	84.19 / 50.90	**86.43 / 59.86**
		Char. Acc. / Word Acc. (Nougat) (%) ↑	22.89 / 16.20	43.11 / 38.00	45.53 / 46.94	4.55 / 11.20	42.34 / 37.15	**53.40 / 48.96**
Deshadowing	RDD	PSNR / SSIM ↑	32.02 / 0.9223	27.88 / 0.7933	32.82 / 0.9281	- / -	**34.35** / 0.9528	33.31 / **0.9532**
		Char. Acc. / Word Acc. (Paddle) (%) ↑	53.65 / 44.31	47.85 / 40.48	57.32 / 47.41	- / -	64.42 / 53.79	**65.69 / 56.23**
		Char. Acc. / Word Acc. (Nougat) (%) ↑	83.59 / 84.13	74.67 / 76.06	83.61 / 83.88	- / -	85.54 / **90.21**	**87.85** / 89.35
		Average rank ↓	4.31	4.17	2.17	5.96	3.06	**1.33**

encoding to each image with a randomly selected quality factor $q \in [2, 12]$, we simulated compression distortions ranging from mild to severe. This process generated a rich set of degradation-clean pairs to train the model for robust recovery across varying compression intensities. For testing, we constructed the TDD_{CAR} set with $q \in [2, 12]$. Additionally, to evaluate robustness under severe compression scenarios, we generated $BCSD_{CAR}$ and $FUNSD_{CAR}$ using a restricted range of $q \in [2, 5]$, as compression artifacts become less challenging at higher quality factors.

Deshadowing. The FSDSRD dataset [30] and the RDD training dataset [29] were combined as the training set for our deshadowing task. FSDSRD consists of 14,200 synthetic images, while the RDD training dataset consists of 4,371 real-world images. The RDD test dataset consisting of 545 real-world images was employed directly as our test dataset. To evaluate the generalization performance on cross-domain data, we further tested on OSR [46] and Jung et al.'s dataset [47], which contains 237 synthetic images and 87 real-world images, respectively.

GT structure maps. For the FSDSRD dataset, the provided binary structure annotations were employed as GT structure maps. For other datasets, GT structure maps were generated by applying Sauvola's adaptive thresholding binarization [48] to GT images.

B. Implementation details

To comprehensively evaluate the performance of DocPure, we conducted extensive benchmarking against a suite of representative SOTA methods, including general-task restoration models (DE-GAN [31], DocDiff [12], and DocStain [33]), and all-in-one restoration models (DocNLC [14] and DocRes [13]). To ensure a fair comparison, all compared models were retrained using their official codes and default configurations. General-task models were trained on individual task-specific training datasets according to restoration tasks: TDD (deblurring), TDD_{noise} (denoising), TDD_{CAR} (CAR), FSDSRD and RDD (deshadowing). All-in-one models (including our DocPure) were trained across all training datasets. Consequently, multiple models were trained for each general-task method, while only one model was trained for each all-in-one method. All compared models were trained on NVIDIA GeForce RTX 4090 GPUs.

The entire training process of the proposed DocPure was decoupled into two stages: the DASAE network was trained first, followed by the restoration network.

Training of our DASAE network. We employed the AdamW optimizer [49] for parameter optimization, with hyperparameters set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate was maintained constant at 1×10^{-4} throughout the training, with a batch size of 32. For the loss function, we set the weights to $\lambda_{bce} = 0.3$, $\lambda_{dice} = 0.5$, and $\lambda_{bd} = 0.35$. In this paper, we have implemented our method with four tasks, *i.e.,* deblurring, denoising, CAR, and deshadowing. Since each sub-encoder is designed to correspond to a restoration task, we set the number of sub-encoders to $k = 4$. In order to improve the training efficiency, we employed a progressive training strategy. First, we conducted pre-training on the deblurring task for 20 epochs by replacing the multi-branch degradation encoder with a single encoder. Then, we froze the structure decoder to serve as a shared structural prediction constraint and proceeded to pre-train new encoders on the denoising, CAR, and deshadowing tasks for 20 epochs, respectively. Finally, we injected the above four pre-trained encoders into the sub-encoders of our multi-branch degradation encoder, unfroze the structure decoder, and trained across all training datasets for 50 epochs. During this training phase, input images for all tasks were randomly cropped to a size of 300×300. The training process took about 60 hours.

Training of our restoration network. The architecture of the restoration network employs a 4-level encoder-decoder structure, with a varying number of Transformer blocks at each level, specifically [4, 6, 6, 8] from level 1 to level 4. We employed a batch size of 16 and trained for a total of 80 epochs. We utilized the AdamW optimizer, with hyperparameters set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate was initialized at 2×10^{-4} and gradually decayed to 1×10^{-6} via the cosine annealing strategy [50]. The learning rate was adjusted at each iteration step to ensure a smoother optimization trajectory. During this phase, images from all tasks were randomly cropped into patches of size 128×128. The training process took about 100 hours.

Fig. 6. Qualitative comparison across four restoration tasks: deblurring, denoising, CAR, and deshadowing, evaluated on the TDD, TDD$_{CAR}$, TDD$_{noise}$, and RDD datasets, respectively. As evidenced by the results on TDD and TDD$_{CAR}$, our method recovers clearer text boundaries and rectifies distorted strokes. This enhanced structural fidelity is consistent with the role of the DASAE, which predicts structure maps and injects high-frequency priors to guide the reconstruction. For denoising on TDD$_{noise}$ and deshadowing on RDD, DocPure achieves clean background separation while preserving intricate details.

C. Metrics

We employed the peak signal-to-noise ratio (PSNR), the structural similarity index (SSIM), and character accuracy as primary quantitative metrics to evaluate the model's performance in document image restoration tasks. PSNR measures the pixel-level reconstruction accuracy between the restored and GT images. SSIM comprehensively assesses the structural similarity between the restored and GT images. Let s_{res} and s_{gt} be the text sequences extracted from the restored and GT images, the OCR character accuracy is defined as $1 - d_c(s_{res}, s_{gt})/\max(|s_{res}|, |s_{gt}|)$, where $|\cdot|$ denotes the sequence length, and $d_c(s_{res}, s_{gt})$ is the character-level edit distance measuring the difference between s_{res} and s_{gt}. Similarly, by tokenizing the texts into word sequences w_{res} and w_{gt}, the OCR word accuracy is computed as $1 - d_w(w_{res}, w_{gt})/\max(|w_{res}|, |w_{gt}|)$, where $d_w(w_{res}, w_{gt})$ is the word-level edit distance between w_{res} and w_{gt}. Two OCR models (Paddle [51] and Nougat [52]) were tested to extract text sequences in our experiments.

D. Experimental results

1) Comparison with SOTA methods: Table II presents a comprehensive quantitative comparison of DocPure against state-of-the-art methods across four restoration tasks. In the deblurring and CAR tasks, our model achieves the best performance among the compared methods, attaining the highest PSNR and SSIM scores, as well as the highest character- and word-level accuracies evaluated using both PaddleOCR and Nougat. In the denoising task, our model exhibits the second-best performance for the six metrics. Regarding the real-world deshadowing task, while DocRes shows a marginal advantage in PSNR and Nougat word accuracy, our model surpasses it in structural fidelity and text recognizability, achieving the highest values of SSIM (0.9532), PaddleOCR character accuracy (65.69%), PaddleOCR word accuracy (56.23%), and Nougat character accuracy (87.85%).

Fig. 6 provides a visual comparison of our method with existing approaches. As shown in the deblurring task (first row), DocPure reconstructs text with clearer structures and continuous strokes compared to other methods. In the denois-

TABLE III

QUANTITATIVE GENERALIZATION COMPARISON WITH SOTA DOCUMENT RESTORATION METHODS ON VARIOUS DEGRADATION DATASETS. GENERALIZATION COMPARISON WAS EVALUATED ON TEST DATASETS WITH STYLES DISTINCT FROM THE TRAINING DATASETS. OUR DOCPURE ACHIEVES THE BEST OVERALL PERFORMANCE. **BOLD** INDICATES THE BEST AND UNDERLINED THE SECOND BEST.

Task	Dataset	Metric	DE-GAN [31] TPAMI'20	DocDiff [12] ACM MM'23	DocStain [33] WACV'25	DocNLC [14] AAAI'24	DocRes [13] CVPR'24	DocPure Ours
Deblurring	FUNSD_{blur}	PSNR / SSIM ↑	25.74 / 0.9666	19.96 / 0.9083	**34.11 / 0.9941**	18.28 / 0.8683	29.22 / 0.9852	31.21 / 0.9899
		Char. Acc. / Word Acc. (Paddle) (%) ↑	73.24 / 39.66	30.66 / 13.50	**91.93 / 67.46**	15.95 / 2.03	81.65 / 59.23	89.79 / 65.45
		Char. Acc. / Word Acc. (Nougat) (%) ↑	51.68 / 49.58	17.65 / 14.69	**76.62 / 78.78**	15.29 / 4.57	66.92 / 67.94	72.90 / 75.03
	BCSD_{blur}	PSNR / SSIM ↑	23.65 / 0.8293	19.78 / 0.6586	28.11 / 0.8637	16.43 / 0.7186	29.57 / 0.8800	**30.33 / 0.8983**
		Char. Acc. / Word Acc. (Paddle) (%) ↑	78.61 / 47.70	67.25 / 24.11	85.14 / 52.40	57.11 / 19.11	85.58 / 50.45	**87.80 / 54.25**
		Char. Acc. / Word Acc. (Nougat) (%) ↑	82.26 / 87.19	81.59 / 81.80	90.38 / 90.78	64.64 / 80.72	90.93 / 88.92	**91.86 / 92.32**
Denoising	FUNSD_{noise}	PSNR / SSIM ↑	39.92 / 0.9955	35.04 / 0.9424	**43.60 / 0.9982**	18.37 / 0.8606	32.20 / 0.9837	41.88 / 0.9955
		Char. Acc. / Word Acc. (Paddle) (%) ↑	96.30 / 79.79	95.09 / 76.02	**97.58 / 87.69**	4.37 / 0.86	95.92 / 72.61	96.73 / 83.53
		Char. Acc. / Word Acc. (Nougat) (%) ↑	87.70 / 83.45	87.17 / 84.33	**90.58** / 86.60	18.19 / 3.79	88.91 / 82.17	89.39 / **87.13**
	BCSD_{noise}	PSNR / SSIM ↑	30.08 / 0.9320	30.57 / 0.8396	32.28 / 0.9376	16.31 / 0.7267	24.51 / 0.5752	**35.15 / 0.9485**
		Char. Acc. / Word Acc. (Paddle) (%) ↑	84.48 / **51.50**	83.26 / 44.64	86.37 / 50.95	62.00 / 23.44	78.35 / 49.45	**86.68** / 50.39
		Char. Acc. / Word Acc. (Nougat) (%) ↑	85.74 / 87.98	89.60 / 89.66	90.28 / 90.38	67.07 / 67.01	86.22 / 90.79	**91.75 / 91.94**
CAR	FUNSD_{CAR}	PSNR / SSIM ↑	21.82 / 0.9484	25.89 / 0.9665	**28.10** / 0.9745	18.39 / 0.8608	26.62 / 0.9727	27.44 / **0.9758**
		Char. Acc. / Word Acc. (Paddle) (%) ↑	63.93 / 24.43	73.98 / 32.24	77.25 / **47.52**	3.39 / 0.73	73.54 / 31.89	**78.72** / 42.48
		Char. Acc. / Word Acc. (Nougat) (%) ↑	41.59 / 33.76	43.96 / 36.37	**61.33** / 40.58	16.51 / 3.28	45.18 / 36.48	57.96 / **48.24**
	BCSD_{CAR}	PSNR / SSIM ↑	17.40 / 0.7297	22.62 / 0.7512	24.85 / 0.7851	16.67 / 0.7189	23.89 / 0.7694	**25.59 / 0.7971**
		Char. Acc. / Word Acc. (Paddle) (%) ↑	74.83 / 43.01	77.46 / 42.27	80.21 / 36.05	54.64 / 18.58	78.54 / 41.28	**80.93 / 46.69**
		Char. Acc. / Word Acc. (Nougat) (%) ↑	82.51 / 82.44	87.39 / 87.42	88.43 / 87.27	79.50 / 79.71	88.32 / 88.33	**88.59 / 88.59**
Deshadowing	OSR	PSNR / SSIM ↑	28.16 / 0.8944	28.30 / 0.8544	**28.66** / 0.9135	- / -	28.53 / 0.9304	**28.66 / 0.9335**
		Char. Acc. / Word Acc. (Paddle) (%) ↑	81.84 / 67.28	73.12 / 55.73	81.33 / 68.45	- / -	81.89 / **72.92**	**82.22** / 70.93
		Char. Acc. / Word Acc. (Nougat) (%) ↑	86.79 / 87.52	85.50 / 86.77	**90.16** / 87.17	- / -	87.27 / 87.44	87.73 / **88.35**
	Jung et al.'	PSNR / SSIM ↑	28.32 / 0.8921	27.69 / 0.8435	27.96 / 0.9047	- / -	28.25 / 0.9082	**28.35 / 0.9090**
		Char. Acc. / Word Acc. (Paddle) (%) ↑	40.17 / 20.81	37.60 / 19.42	40.27 / 20.25	- / -	40.04 / 20.81	**40.33 / 20.95**
		Char. Acc. / Word Acc. (Nougat) (%) ↑	35.87 / 37.57	32.23 / 21.92	35.92 / 36.49	- / -	34.02 / 37.83	**36.73 / 38.06**
		Average rank ↓	3.85	4.44	2.16	5.94	3.20	**1.40**

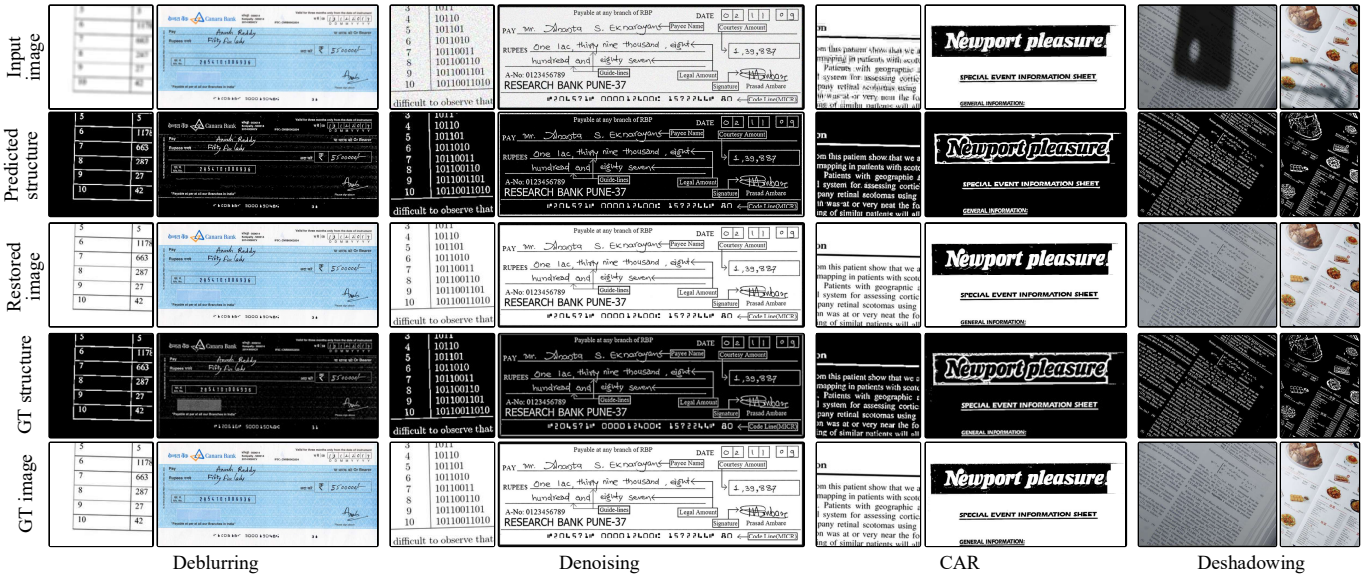

Fig. 7. Visual effects of our DocPure on diverse datasets with unseen styles (from top to bottom): degraded input images, our predicted structural maps, our restored images, GT structure maps, and GT images. The visual results show that our method recovers document content and structural details, suggesting generalization capabilities across different document styles and degradation types.

ing scenario (second row), our method robustly recovers clean details from Gaussian noise corruption, avoiding the over-smoothing or residual noise commonly observed in baselines. For the CAR task (third row), the model handles complex high-frequency compression noise, reducing blocking artifacts while maintaining the integrity of text strokes. Qualitative results for deshadowing (fourth row) demonstrate that DocPure restores uniform illumination and maintains stable font morphology, mitigating the background residues and semantic blurring prevalent in existing all-in-one models. Since DocNLC requires cross-task pairing for the training data, it cannot be trained with the unpaired deshadowing dataset.

The consistently strong performance across these diverse tasks suggests the effectiveness of our unified framework, rather than relying on task-specific engineering. The successful recovery of fine-grained structure details supports the effectiveness of DASAE in predicting clear structural priors from corrupted inputs. The robust performance across multiple tasks underscores the efficacy of leveraging the distinctiveness of degradations in the wavelet domain and endowing the

TABLE IV
COMPARISON OF MODEL PROPERTIES AMONG DIFFERENT DOCUMENT RESTORATION METHODS. OUR DOCPURE IS THE ONLY MODEL AMONG THE COMPARED METHODS SUPPORTING ALL THREE PROPERTIES: ALL-IN-ONE INFERENCE, PROMPT-FREE INFERENCE, AND CROSS-TASK PAIRING-FREE TRAINING DATA.

Model property	DE-GAN [31]	DocDiff [12]	DocStain [33]	DocNLC [14]	DocRes [13]	DocPure (ours)
All-in-one inference	✗	✗	✗	✓	✓	✓
Prompt-free inference	✗	✗	✗	✓	✗	✓
Cross-task pairing-free	✓	✓	✓	✗	✓	✓

TABLE V
STATISTICS OF NUMBER OF PARAMETERS (#PARAMS) AND COMPUTATIONAL COMPLEXITY (GFLOPS) AT THE RESOLUTION OF 256×256.

Metric	DE-GAN [31]	DocDiff [12]	DocStain [33]
#Params (M)	30.00	8.20	6.86
GFLOPs	109.00	4876.50	283.90
Metric	DocNLC [14]	DocRes [13]	DocPure (ours)
#Params (M)	7.80	15.20	38.15
GFLOPs	34.80	183.00	446.82

TABLE VI
STATISTICS OF EFFICIENCY UNDER DIFFERENT NUMBER OF SUB-ENCODERS k AT THE RESOLUTION OF 1024×1024.

k value	DASAE			Full model (DocPure)		
	Params	Latency	GPU memory	Params	Latency	GPU memory
$k = 1$	4.06M	55.5ms	2.194GB	34.60M	187.0ms	4.929GB
$k = 2$	5.23M	78.7ms	2.209GB	35.81M	209.6ms	4.936GB
$k = 3$	6.40M	100.0ms	2.491GB	36.98M	231.9ms	4.942GB
$k = 4$	7.57M	122.4ms	3.014GB	38.15M	253.9ms	4.948GB

model with degradation awareness through wavelet-spatial interaction.

2) Generalization ability assessment: To validate the model's robustness, we evaluated the cross-domain generalization capability of DocPure on document styles and layouts that were unseen during training. As shown in Table III, our method consistently maintains top-tier performance across unseen datasets. For example, in the deshadowing task, DocPure achieves the best performance on the OSR dataset in terms of PSNR (28.66 dB), SSIM (0.9335), and the PaddleOCR character accuracy (82.22%), and the second-best Nougat character accuracy (87.73%); On the real-world Jung et al. datasets, DocPure further achieves the best performance among the compared methods for all six evaluation metrics, highlighting the effectiveness of our unified architecture. DocPure's performance on challenging unseen document styles and layouts attests to the model's strong feature adaptability.

Fig. 7 visually demonstrates the model's robustness against unseen styles. Even when handling document layouts with complex fonts or variable spacing, our method reconstructs characters with sharp boundaries and continuous strokes, avoiding the fractured text artifacts common in other methods. In contrast to some compared methods that suffer from semantic blurring, DocPure maintains morphological consistency, validating its degradation-aware generalization.

3) Discussion on model properties: As illustrated in Table IV, existing document image restoration methods struggle to simultaneously satisfy the requirements of unified modeling, prompt-free inference, and cross-task pairing-free training data. First of all, previous general-task document restoration models, such as DE-GAN, DocDiff, and DocStain, require training separate weights for different degradation types (*i.e.,* several models for several degradations), which scales poorly in real-world scenarios. On the contrary, our unified framework enables adaptive high-quality document restoration by taking advantage of the proposed degradation-informed routing regularization mechanism. Second, DocRes achieves unified modeling but relies heavily on explicit task prompts at inference to specify the degradation category. Consequently, it fails to perform degradation-aware restoration when the degradation prior is unknown at inference. In contrast, our DocPure eliminates this dependency by leveraging the proposed DASAE to predict the structure map to guide the prompt-free degradation-aware inference. Third, although DocNLC achieves prompt-free unified restoration, its contrastive learning framework imposes a stringent constraint on dataset construction. Specifically, it requires strictly paired multi-degradation data (*i.e.,* multiple distinct degradation types must correspond to the exact same clean ground truth), which is highly restrictive and costly to synthesize or collect at scale. Fortunately, our DocPure overcomes this limitation by operating effectively without such rigorous pairing demands. In conclusion, DocPure stands out as the only method that achieves prompt-free document restoration with a single unified model while avoiding the strict data pairing constraint. This makes our approach more practical for diverse document restoration settings.

4) Analysis on computational complexity: Table V provides statistics of the number of parameters and computational complexity (GFLOPs). From the table, we can see that DocDiff suffers from massive computational overhead (4876.50 GFLOPs) due to its iterative generation process. DocNLC has fewer parameters and fewer computations, but its restoration quality in terms of PSNR and downstream character accuracy falls considerably across multiple degradation scenarios. While DocPure is substantially lighter than the diffusion-based DocDiff (446.82 GFLOPs vs. 4876.50 GFLOPs), it is heavier than other non-diffusion all-in-one baselines such as (34.80 GFLOPs) and DocRes (183.00 GFLOPs). This reflects that the design choice of DocPure prioritizes restoration quality and degradation awareness over computational minimalism.

Table VI reports statistics of the number of parameters, inference latency, and peak GPU memory usage for both DASAE and the full model under different numbers of sub-encoders k at the resolution of 1024×1024. With the increase of k, the total computational overhead of our multi-branch DocPure increases moderately (around 20ms) compared to the

TABLE VII
ABLATION STUDY OF THE PROPOSED COMPONENTS ON VARIOUS DOCUMENT RESTORATION TASKS. OUR DOCPURE ACHIEVES THE BEST OVERALL PERFORMANCE. THE BEST PERFORMANCE IS HIGHLIGHTED IN **BOLD**.

Configuration	DASAE	SWIM	Deblurring		Denoising		CAR		Deshadowing	
			PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Baseline	✗	✗	28.80	0.9838	40.13	0.9960	29.68	0.9850	32.01	0.9513
w/o DASAE	✗	✓	31.66	0.9901	40.70	0.9968	30.88	0.9885	32.64	0.9521
w/o SWIM	✓	✗	29.84	0.9861	40.15	0.9960	30.04	0.9861	32.22	0.9517
Ours	✓	✓	**32.55**	**0.9912**	**41.26**	**0.9973**	**31.34**	**0.9893**	**33.31**	**0.9532**

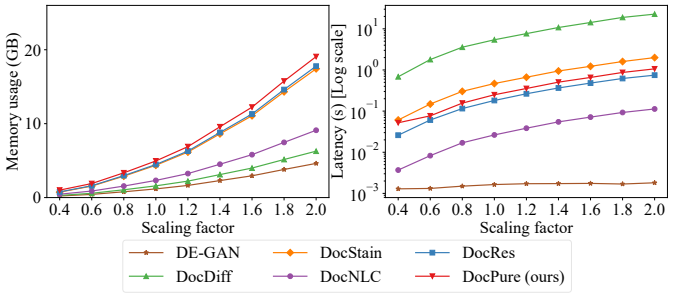

Fig. 8. Peak GPU memory usage (left) and average inference latency (right) of compared methods at various scaling factors applied to a 1024 × 1024 resolution image.

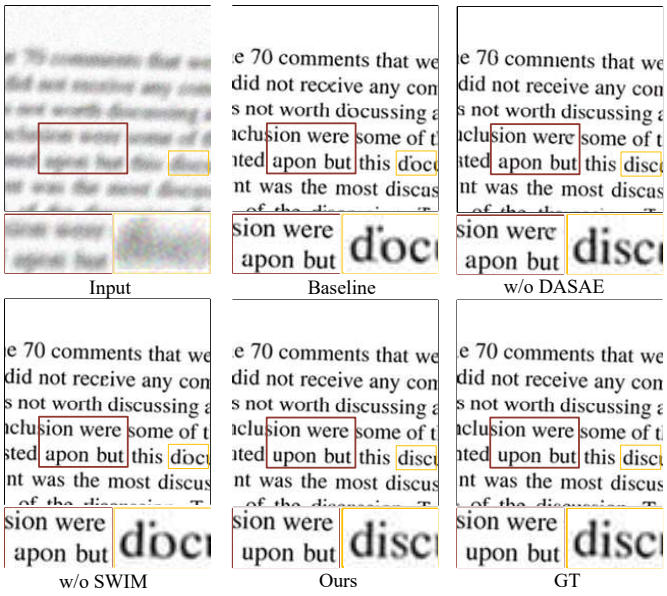

Fig. 9. Visual ablation comparison of different components. The baseline and variants without specific modules (w/o DASAE, w/o SWIM) fail to recover semantically correct text, resulting in character substitution errors. In contrast, our full model integrates all proposed modules to restore sharp, legible, and semantically accurate text, matching the GT.

inference latency (187ms) of the single-branch model ($k = 1$). The structure predictor is required for every degraded input. The results exhibit a good trade-off between the performance improvement and the computational overhead achieved by our multi-branch sub-encoder architecture. Our multi-branch design is scalable. A sub-encoder can be added to our DASAE for a new degradation type. This requires training the new sub-encoder, incrementally fine-tuning DASAE, and adapting the restoration backbone with the updated structure predictor. Based on the one-to-one correspondence, the number of sub-encoders grows linearly with the number of degradation types. As shown in Table VI, the number of parameters of the full models increases slightly with nearly stable peak GPU memory usage.

In Fig. 8, we visualize the maximum GPU memory usage and inference latency for the compared models when processing images of various resolutions. Our method exhibits slightly higher memory usage compared to the DocRes and DocStain baselines. Although the multi-branch architecture suggests higher computational costs, we can see that our method still maintains a moderate computational overhead. For example, at a resolution of 1024 × 1024, the inference latency of our DocPure is 253.9ms per image. While this is slower than DocRes (181.3ms), it is significantly faster than DocDiff (5413.9ms) and DocStain (467.6ms).

E. Ablation study

Table VII quantitatively validates the contributions of the predicted structure map and the wavelet interaction with the proposed DASAE and SWIM modules, respectively. First, the w/o DASAE variant fails to achieve the optimal performance. This performance drop confirms that without the explicit structural priors learned by the DASAE, the restoration network lacks precise structure guidance, leading to suboptimal stroke reconstruction. Second, we observe a distinct performance degradation in the w/o SWIM configuration. This quantitative gap provides evidence that relying solely on spatial convolutions is less effective for handling complex frequency-dependent degradations. The wavelet-domain interaction is important for recovering high-frequency structure details. The best overall performance of the full DocPure framework demonstrates the synergistic effect of combining degradation-aware structural extraction with frequency-guided modulation.

Fig. 9 illustrates the visual results in the absence of specific modules. In the absence of structure maps predicted by DASAE serving as internal priors, the restoration network loses structural constraints. Without such structure maps, the model struggles to preserve character strokes and capture complex text structures. Upon the removal of the SWIM module, the model loses degradation-aware spectral guidance. In this scenario, the model tends to project multiple heterogeneous degradation patterns into a unified feature space for processing, leading to severe cross-task interference. Due to the lack of targeted guidance and enhancement for high-frequency components, the restored text structures are not sufficiently sharp and may also exhibit noticeable ghosting or artifacts.

TABLE VIII

SENSITIVITY ANALYSIS OF LOSS WEIGHTS λ_{bce}, λ_{dice}, AND λ_{bd}. FOR EACH PARAMETER GROUP, ONLY THE CORRESPONDING WEIGHT IS VARIED, WHILE THE OTHER TWO ARE FIXED TO THE DEFAULT CONFIGURATION $(\lambda_{bce}, \lambda_{dice}, \lambda_{bd}) = (0.30, 0.50, 0.35)$. THE BEST PERFORMANCE IN EACH PARAMETER GROUP IS HIGHLIGHTED IN **BOLD**.

Parameter	Value	Deblurring			Denoising			CAR			Deshadowing		
		PSNR↑	SSIM↑	Char. Acc.↑	PSNR↑	SSIM↑	Char. Acc.↑	PSNR↑	SSIM↑	Char. Acc.↑	PSNR↑	SSIM↑	Char. Acc.↑
λ_{bce}	0.00	32.44	0.9908	55.84	41.20	0.9971	72.29	31.23	0.9890	53.09	33.22	0.9526	87.31
	0.15	32.48	0.9911	56.13	**41.27**	**0.9973**	**72.33**	31.32	0.9891	53.29	33.30	0.9532	87.76
	0.30	32.55	**0.9912**	56.31	41.26	**0.9973**	72.30	**31.34**	**0.9893**	**53.40**	**33.31**	0.9532	**87.85**
	0.45	**32.56**	**0.9912**	**56.34**	41.24	**0.9973**	72.23	31.26	0.9891	53.15	**33.31**	**0.9533**	87.84
	0.60	32.52	0.9910	55.93	41.24	**0.9973**	72.22	31.27	0.9891	53.13	33.29	0.9531	87.73
λ_{dice}	0.00	32.42	0.9909	55.78	41.22	**0.9973**	72.22	31.21	0.9888	52.97	33.16	0.9524	87.04
	0.25	32.50	0.9911	56.07	41.24	**0.9973**	72.26	31.28	0.9891	53.22	33.28	0.9530	87.77
	0.50	32.55	0.9912	56.31	**41.26**	**0.9973**	72.30	31.34	**0.9893**	**53.40**	**33.31**	**0.9532**	**87.85**
	0.75	**32.59**	**0.9913**	**56.42**	**41.26**	**0.9973**	**72.33**	**31.36**	**0.9893**	53.37	33.27	0.9528	87.82
	1.00	32.49	0.9910	56.06	41.25	**0.9973**	72.29	31.31	0.9892	53.28	33.28	0.9527	87.66
λ_{bd}	0.00	32.45	0.9909	55.88	41.24	**0.9973**	72.26	31.25	0.9891	53.12	33.22	0.9527	87.33
	0.10	32.47	0.9910	55.90	41.24	**0.9973**	72.23	31.27	0.9891	53.14	33.24	0.9527	87.36
	0.35	**32.55**	**0.9912**	**56.31**	**41.26**	**0.9973**	**72.30**	**31.34**	**0.9893**	**53.40**	33.31	**0.9532**	87.85
	0.50	**32.55**	0.9911	56.29	**41.26**	**0.9973**	72.28	31.28	0.9892	53.25	**33.33**	**0.9532**	**87.91**
	0.70	32.51	0.9910	56.03	41.25	**0.9973**	72.27	31.27	0.9892	53.22	33.31	**0.9532**	87.84

TABLE IX

COMPARISON OF UTILIZING DIFFERENT WAVELET BASES FOR FREQUENCY-DOMAIN DECOMPOSITION.

Wavelet basis	Deblurring		Denoising		CAR		Deshadowing	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Haar [15]	**32.55**	**0.9912**	**41.26**	**0.9973**	31.34	**0.9893**	33.31	0.9532
Db2 [53]	32.39	0.9898	41.21	0.9966	**31.51**	0.9889	**33.46**	**0.9539**

TABLE X

ABLATION STUDY ON THE DESIGN CHOICES OF DASAE AND SWIM WITH SIMPLER ALTERNATIVES: (A) A SINGLE SHARED ENCODER WITH TASK-LABEL AUXILIARY LOSS REPLACING THE MULTI-BRANCH DEGRADATION ENCODER IN DASAE; (B) STANDARD WAVELET CONCATENATION WITHOUT CROSS-ATTENTION; (C) SOBEL EDGE MASK PRIORS INSTEAD OF LEARNED STRUCTURE MAPS; (D) FILM MODULATION WITHOUT THE FULL CFAM DESIGN; (E) A DEGRADATION CLASSIFIER PLUS CONDITIONAL NORMALIZATION.

Setting	Deblurring		Denoising		CAR		Deshadowing	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
A	32.38	0.9905	41.22	0.9972	31.27	0.9890	33.06	0.9522
B	30.52	0.9865	40.58	0.9956	30.13	0.9864	32.33	0.9521
C	31.21	0.9891	41.17	0.9965	31.18	0.9871	32.72	0.9516
D	31.97	0.9905	40.94	0.9966	31.20	0.9868	32.69	0.9521
E	30.93	0.9887	40.59	0.9962	30.62	0.9878	32.42	0.9528
Ours	**32.55**	**0.9912**	**41.26**	**0.9973**	**31.34**	**0.9893**	**33.31**	**0.9532**

We conducted a sensitivity analysis of the loss weights λ_{bce}, λ_{dice}, and λ_{bd}. As reported in Table VIII, the overall performance varies slightly across different weight settings, showing that our DocPure is not overly sensitive to a particular weight choice. The default setting (λ_{bce}, λ_{dice}, λ_{bd}) = (0.30, 0.50, 0.35) achieves the best overall performance across the four document restoration tasks. The BCE loss provides stable pixel-level supervision for dense structure-map prediction, the Dice loss encourages region-level consistency between text foreground and background, and the boundary loss improves stroke contour localization.

Table IX provides the comparison of utilizing the Haar wavelet [15] and the Daubechies 2 (db2) wavelet [53] for frequency-domain decomposition within SWIM. The Haar wavelet is suitable for preserving abrupt edges in sharp font strokes, while the db2 wavelet is more suitable for modeling the restoration of low-frequency continuous signals such as shadow gradients. The experimental results show that the performance of the two choices is comparable overall. Haar achieves better performance in deblurring and denoising, while db2 performs better in deshadowing.

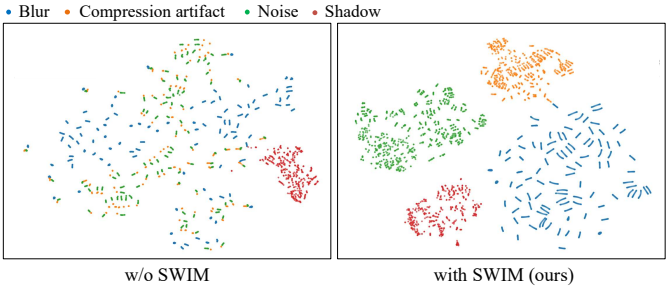

Fig. 10. Comparison of t-SNE visualization of latent feature distributions extracted from the model bottlenecks without and with the SWIM block. The inclusion of SWIM results in well-separated clusters, indicating improved degradation discriminability.

As shown in Table X, we conducted an ablation study to justify the design choices of DASAE and SWIM by comparing them with several baseline variants. The tested alternatives include: (A) a single shared encoder with task-label auxiliary loss, (B) standard wavelet concatenation without cross-attention, (C) Sobel mask priors instead of learned structure maps, (D) FiLM modulation without the full CFAM design, and (E) a degradation classifier plus conditional normalization. The results demonstrate that the design choices of DASAE and SWIM consistently achieve the best performance across all four restoration tasks. Therefore, this ablation study supports the proposed design choices of degradation-aware structure prediction and structure-guided frequency modulation for unified document restoration.

Fig. 10 visualizes the comparison of latent feature distributions extracted from the models with and without SWIM. Compared to the baseline without SWIM, our model successfully maps different degradations to well-separated clusters. This visualization provides qualitative evidence that SWIM

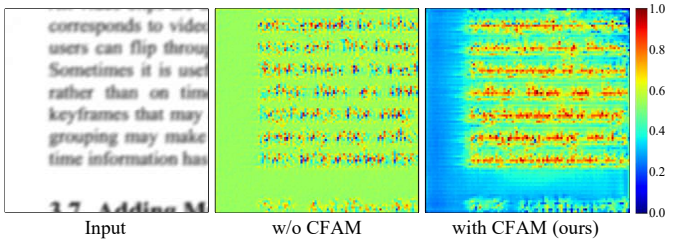

Fig. 11. Feature heatmaps without and with the CFAM module. The lack of the CFAM module hinders the separation of text from background, causing ambiguous feature activations. Conversely, CFAM enhances localization capability, guiding the model to focus on informative text features.

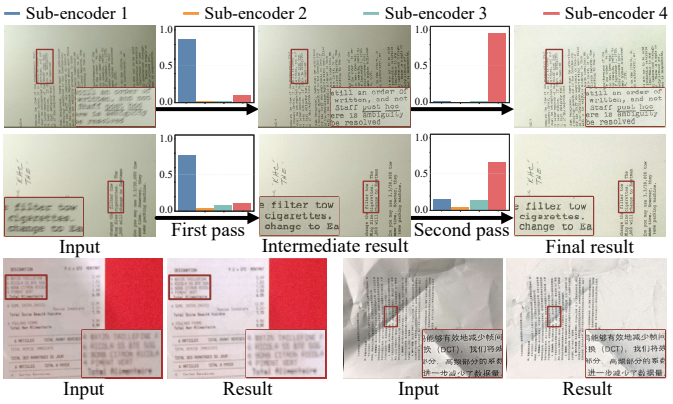

Fig. 12. Examples of our restoration results under real-world mixed degradation scenarios. Our method can mitigate real-world degradations in documents with moderate blur and shadow (top), while tending to fail to restore extremely blurred small fonts or paper-crease degradations unseen during training (bottom).

improves degradation-discriminative representations, supporting input-adaptive restoration without manual task prompts.

To visually demonstrate the effectiveness of CFAM, Fig. 11 presents the feature heatmaps with and without the integration of CFAM. In the configuration without CFAM, the feature activation distribution appears relatively diffuse. Consequently, the model struggles to effectively differentiate between text strokes and background noise, leading to severe ambiguity in feature responses. In contrast, upon introducing CFAM, the low-frequency structural priors help calibrate the high-frequency feature distribution through the dual mechanism of spatial gating and channel modulation. This enables the model to focus more precisely on informative text regions, enhancing the saliency of text structures while suppressing artifacts in non-text regions.

F. Discussion on real-world document restoration

In Table II and Table III, we have demonstrated that our method is capable of restoring documents on a single real-world degradation with the real-world deshadowing datasets (RDD [29] and Jung et al.'s dataset [47]). Here we further test the performance of our method under real-world mixed degradation scenarios, where multiple degradation types coexist in a single image. Since our multi-branch model design focuses on the restoration of a single dominant degradation, we ran multiple inference passes on mixed degradations to obtain the restored result. As shown in Fig. 12 (top), our method can mitigate real-world degradations in documents with moderate blur and shadow. The routing mechanism produces a weighted combination of sub-encoder features, where the dominant degradation is distinguished by the largest weight. The figure shows that our method successfully performs deblurring in the first inference pass, followed by deshadowing in the second. However, our method tends to fail to restore extremely blurred small fonts or paper-crease degradations unseen during training. The failure cases in Fig. 12 (bottom) show that the restored results still contain incomplete text details or local residual degradations. This is mainly because our method is currently designed around a known set of trained degradations. Nevertheless, this experiment indicates that our method has potential for real-world mixed degradation scenarios.

V. Conclusions

We propose DocPure, a prompt-free degradation-aware framework for unified document image restoration. By explicitly identifying degradation types to guide structural priors via degradation-informed routing regularization in the latent space, while implicitly perceiving degradation patterns through collaborative interaction in the wavelet domain, our DocPure achieves an effective balance between the restoration accuracy and generalization capability across diverse document types. The degradation-label supervision is only applied during training, and no manual task prompts or degradation labels are required at inference. The proposed degradation-aware structure auto-encoder and structure-guided wavelet interaction module are integrated into a unified U-shaped architecture. This design enables the model to identify diverse degradation modes via spectral signatures, while leveraging cross-frequency adaptive modulation to guide features toward structure-sharp representations and suppress high-frequency artifacts. Extensive experiments on 12 datasets across four representative document restoration tasks demonstrate that our prompt-free DocPure not only achieves strong performance compared with SOTA methods without requiring task prompts, but also exhibits robustness on unseen document styles.

Limitations and future work. Our method comes with certain limitations. First, as mentioned beforehand, our method is based on a known set of trained degradations. Adapting to a new degradation category may require additional incremental training or an expansion of the routing-space capacity. Second, for mixed degradations, spatially non-uniform degradations, or degradations outside the training degradation set, the model does not have a dedicated routing pathway, and the weighted combination of sub-encoder features may not fully capture the complex degradation distribution. Developing a more scalable blind document restoration framework remains an interesting future direction. Third, the memory consumption becomes high when restoring high-resolution documents, which limits the deployment on resource-limited edge devices. We will investigate a more lightweight model that utilizes techniques such as efficient context modeling [54] and model distillation. In addition, we would like to extend our method to general image restoration applications in the near future.

SUPPLEMENTARY DOCUMENT

I. More implementation details

Algorithm 1 Pseudocode of DocPure

Predict structure map $\mathbf{I}_M \leftarrow \text{DASAE}(\mathbf{I})$
Extract shallow features $\mathbf{Y}_0 \leftarrow \text{Conv}(\mathbf{I})$
for $l = 1$ to 3 **do** # Encoding phase
  $\mathbf{Y}'_l \leftarrow \text{TransformerBlocks}(\mathbf{Y}_{l-1})$
  $\mathbf{E}_l \leftarrow \mathbf{Y}'_l$ # Save for skip connection
  $\mathbf{Y}_l \leftarrow \text{PixelUnshuffle}(\text{Conv}(\mathbf{Y}'_l))$ # Downsample
end for
$\mathbf{Y}_4 \leftarrow \text{TransformerBlocks}(\mathbf{Y}_3)$ # Bottleneck layer
for $l = 3$ down to 1 **do** # Decoding phase
  $\mathbf{Y}_{up} \leftarrow \text{PixelShuffle}(\text{Conv}(\mathbf{Y}_{l+1}))$ # Upsample
  $\mathbf{Y}' \leftarrow \text{SWIM}(\mathbf{Y}_{up}, \mathbf{I}, \mathbf{I}_M)$
  $\mathbf{Y}'' \leftarrow [\mathbf{Y}', \mathbf{E}_l]$ # Skip connection
  if $l > 1$ **then**
    $\mathbf{Y}_l \leftarrow \text{TransformerBlocks}(\mathbf{Y}'')$
  end if
end for
$\hat{\mathbf{I}} \leftarrow \text{TransformerBlocks}(\mathbf{Y}'')$
return $\hat{\mathbf{I}}$ # Restored document image

Algorithm 2 Pseudocode of DASAE

# Multi-branch disentanglement encoding
for $i = 1$ to k **do**
  $\mathbf{X}^0_{down,i} \leftarrow \mathbf{I}$
  for $m = 1$ to 3 **do**
    $\mathbf{X}^m_i \leftarrow \text{Conv}_{3\times3}(\text{Conv}_{3\times3}(\mathbf{X}^{m-1}_{down,i}))$
    $\mathbf{X}^m_{down,i} \leftarrow \text{MaxPool}(\mathbf{X}^m_i)$ # Downsample
  end for
  $\mathbf{X}^4_i \leftarrow \mathbf{X}^3_{down,i}$ # Deepest feature representation
end for
# Degradation perception gate (DPG)
$[w_1, \cdots, w_k] \leftarrow \text{Softmax}(\mathcal{SE}(\mathbf{X}^4_1, \ldots, \mathbf{X}^4_k))$
# Feature aggregation
for $m = 1$ to 3 **do**
  $\mathbf{X}^m_{skip} \leftarrow \sum_{i=1}^k w_i \cdot \mathbf{X}^m_i$ # Adaptive skip features
end for
$\mathbf{X}_{task} \leftarrow \sum_{i=1}^k w_i \cdot \mathbf{X}^4_i$
$\mathbf{X}_{ASPP} \leftarrow \text{ASPP}(\mathbf{X}_{task})$
$\mathbf{X}^3_{in} \leftarrow \mathbf{X}_{ASPP}$
# Structure decoding phase
for $m = 3$ to 1 **do**
  $\mathbf{X}^m_{up} \leftarrow \text{ConvTranspose}(\mathbf{X}^m_{in})$ # Upsample
  # Feature fusion
  $\mathbf{X}^m_{dec} \leftarrow \text{ConvBNReLU}(\text{ConvBNReLU}(\left[\mathbf{X}^m_{up}, \mathbf{X}^m_{skip}\right]))$
  if $m > 1$ **then**
    $\mathbf{X}^{m-1}_{in} \leftarrow \mathbf{X}^m_{dec}$
  end if
end for
$\mathbf{I}_M \leftarrow \text{Sigmoid}(\text{Conv}_{1\times1}(\mathbf{X}^1_{dec}))$
return $\mathbf{I}_M$ # Predicted structure map

Algorithm 3 Pseudocode of SWIM

Wavelet decomposition & feature mapping
$(\mathbf{I}_{LL}, \mathbf{I}_{HL}, \mathbf{I}_{LH}, \mathbf{I}_{HH}) \leftarrow \mathcal{W}_{Haar}(\mathbf{I})$
$(\mathbf{M}_{LL}, \mathbf{M}_{HL}, \mathbf{M}_{LH}, \mathbf{M}_{HH}) \leftarrow \mathcal{W}_{Haar}(\mathbf{I}_M)$
$\mathbf{F}_L \leftarrow \text{Conv}_{3\times3}(\mathbf{I}_{LL})$
$\mathbf{F}^{rgb}_H \leftarrow \text{Conv}_{3\times3}([\mathbf{I}_{HL}, \mathbf{I}_{LH}, \mathbf{I}_{HH}])$
$\mathbf{F}^{st}_H \leftarrow \text{Conv}_{3\times3}([\mathbf{M}_{HL}, \mathbf{M}_{LH}, \mathbf{M}_{HH}])$
Frequency semantic aggregation
$\mathbf{F}'_L \leftarrow \text{CA}(\mathbf{Q}=\mathbf{F}_L, \mathbf{K}=\mathbf{Y}, \mathbf{V}=\mathbf{Y})$
$\mathbf{F}'^{rgb}_H \leftarrow \text{CA}(\mathbf{Q}=\mathbf{F}^{rgb}_H, \mathbf{K}=\mathbf{Y}, \mathbf{V}=\mathbf{Y})$
$\mathbf{F}'^{st}_H \leftarrow \text{CA}(\mathbf{Q}=\mathbf{F}^{st}_H, \mathbf{K}=\mathbf{Y}, \mathbf{V}=\mathbf{Y})$
Spatial gating to balance textures and structures
$\mathbf{W}_1, \mathbf{W}_2 \leftarrow \text{G}([\mathbf{F}^{rgb}_H, \mathbf{F}^{st}_H])$
$\mathbf{F}'_H \leftarrow \text{Conv}_{1\times1}(\mathbf{W}_1 \odot \mathbf{F}'^{rgb}_H + \mathbf{W}_2 \odot \mathbf{F}'^{st}_H)$
Begin of cross-frequency adaptive modulation (CFAM)
  # Enhance high-frequency
  $\mathbf{F}''_H \leftarrow \text{DWC}(\text{DWC}(\text{Conv}_{1\times1}(\mathbf{F}'_H)))$
  # Dual modulation
  $\mathbf{F}''_L \leftarrow \text{Conv}_{1\times1}(\mathbf{F}'_L)$
  $\mathbf{G}_{map} \leftarrow \text{Sigmoid}(\text{Conv}_{1\times1}(\text{DBD}(\mathbf{F}''_L)))$
  $\boldsymbol{\gamma}, \boldsymbol{\beta} \leftarrow \text{Conv}_{1\times1}(\text{GAP}(\mathbf{F}''_L))$
  $\tilde{\mathbf{F}}_H \leftarrow \boldsymbol{\gamma} \odot (\mathbf{F}''_H \odot \mathbf{G}_{map}) + \boldsymbol{\beta}$
  $\mathbf{F}_{out} \leftarrow \text{Concatenation}(\tilde{\mathbf{F}}_H, \mathbf{F}'_L)$
End of CFAM
$\mathbf{Y} \leftarrow \text{CA}(\mathbf{Q}=\mathbf{Y}, \mathbf{K}=\mathbf{F}_{out}, \mathbf{V}=\mathbf{F}_{out})$
return $\mathbf{Y}$ # High-frequency enhanced features

Algorithm 1 provides the pseudocode of our prompt-free unified document restoration framework (DocPure). Algorithm 2 outlines the pseudocode of our degradation-aware structure auto-encoder (DASAE). Algorithm 3 describes the pseudocode of our structure-guided wavelet interaction module (SWIM).

II. More experimental results

A. More ablation study

Table XI presents the ablation study of the internal components of DASAE, including the multi-branch encoder, the degradation perception gate (DPG), and the routing regularization (RR). The quantitative results demonstrate that, compared to the single encoder baseline or the variant without DPG employing simple average fusion, the proposed multi-branch encoder equipped with DPG without RR significantly improves the accuracy of structure prediction. Furthermore, with the incorporation of routing regularization, the DPG can assign weights more precisely. This mechanism stabilizes the routing process, thereby further enhancing the efficacy of structure restoration.

Fig. 13 shows a visual comparison of ablation results for DASAE. The single encoder baseline forces the network to learn an entangled distribution of degradations. The blurred artifacts reveal the model's struggle in this entangled space. In contrast, the introduction of routing regularization in DASAE results in sharp structural maps, demonstrating that the degradation-discriminative routing strategy success-

TABLE XI
ABLATION STUDY OF THE INTERNAL COMPONENTS OF DASAE ON VARIOUS DOCUMENT RESTORATION TASKS. THE BEST PERFORMANCE IS HIGHLIGHTED IN **BOLD**.

Configuration	Multi-branch encoder	Degradation perception gate (DPG)	Routing regularization in DPG	Deblurring		Denoising		CAR		Deshadowing	
				PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Single encoder	✗	✗	✗	31.98	0.9907	41.16	0.9972	31.17	0.9889	32.86	0.9525
w/o DPG	✓	✗	✗	31.80	0.9904	41.18	0.9972	31.17	0.9888	32.83	0.9525
w/o RR	✓	✓	✗	32.41	0.9910	41.21	0.9972	31.30	0.9892	33.14	0.9527
DASAE (ours)	✓	✓	✓	**32.55**	**0.9912**	**41.26**	**0.9973**	**31.34**	**0.9893**	**33.31**	**0.9532**

TABLE XII
ABLATION STUDY OF CFAM ON VARIOUS DOCUMENT RESTORATION TASKS. THE RESULTS DEMONSTRATE THE PERFORMANCE GAIN ACROSS ALL FOUR TASKS WHEN CFAM IS EMPLOYED. THE BEST PERFORMANCE IS HIGHLIGHTED IN **BOLD**.

Configuration	Deblurring		Denoising		CAR		Deshadowing	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
w/o CFAM	31.46	0.9898	40.64	0.9967	30.59	0.9875	33.03	0.9528
with CFAM (ours)	**32.55**	**0.9912**	**41.26**	**0.9973**	**31.34**	**0.9893**	**33.31**	**0.9532**

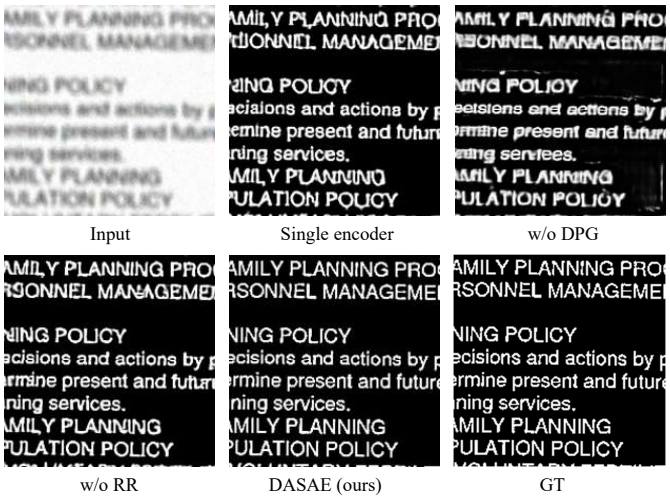

Fig. 13. Visual comparison of ablation results for DASAE. The visualizations correspond to the ablation settings listed in Table XI. Utilizing a single encoder maps all degradation types into an entangled space. Furthermore, the absence of either DPG or routing regularization hinders the extraction of accurate degradation representations with blurry artifacts. In contrast, DASAE restores images with sharp structures.

fully prevents mode collapse and ensures task-specific feature extraction.

Table XII illustrates the performance gains attributed to CFAM. The enhancement is particularly pronounced in the deblurring task, which requires greater fidelity in restoring high-frequency details. This indicates that CFAM plays a crucial role in preserving structural textures. CFAM adaptively modulates the response of high-frequency features via spatial gating and channel modulation. This cross-frequency interaction ensures that the restored high-frequency details are consistently aligned with the low-frequency content in terms of global structure.

B. More visual results

As shown in Fig. 14, to further demonstrate the robustness of our method, we provide additional qualitative comparisons against the state-of-the-art (SOTA) methods on four primary document restoration tasks: deblurring, denoising, compression artifact reduction (CAR), and deshadowing. The compared SOTA methods include general-task restoration models (DE-GAN [31], DocDiff [12], and DocStain [33]), and all-in-one restoration models (DocNLC [14] and DocRes [13]). For the deshadowing task, since DocNLC requires cross-task pairing for the training data, it cannot be trained with the unpaired deshadowing dataset.

Fig. 15 illustrates the robustness of our model against distribution shifts. In contrast to existing SOTA methods, which are prone to structural distortions or residual artifacts in cross-domain scenarios, DocPure exhibits superior generalization capabilities, enabling the high-quality restoration of clear text content.

REFERENCES

[1] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. DocFormer: End-to-end transformer for document understanding. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 973–983, 2021. 1

[2] Rui Shu, Cairong Zhao, Shuyang Feng, Liang Zhu, and Duoqian Miao. Text-enhanced scene image super-resolution via stroke mask and orthogonal attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 33:6317–6330, 2023. 1

[3] Runmin Wang, Shengrong Yuan, Shan Ye, Xingdong Song, Han Xu, Shengyou Qian, Changxin Gao, and Nong Sang. Msinet: A mask structure inference network for scene text image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2026. 1

[4] Yuhong Zhang, Hengsheng Zhang, Xinning Chai, Zhengxue Cheng, Rong Xie, Li Song, and Wenjun Zhang. Diff-restorer: Unleashing visual prompts for diffusion-based universal image restoration. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2025. 1

[5] Naoto Inoue and Toshihiko Yamasaki. Learning from synthetic shadows for shadow detection and removal. *IEEE Transactions on Circuits and Systems for Video Technology*, 31:4187–4197, 2021. 1

[6] Shaokai Liu, Hao Feng, and Wengang Zhou. Rethinking supervision in document unwarping: A self-consistent flow-free approach. *IEEE Transactions on Circuits and Systems for Video Technology*, 34:4817–4828, 2024. 1

[7] Zhi Jin, Yuwei Qiu, Kaihao Zhang, Hongdong Li, and Wenhan Luo. Mb-taylorformer v2: Improved multi-branch linear transformer expanded by taylor formula for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1

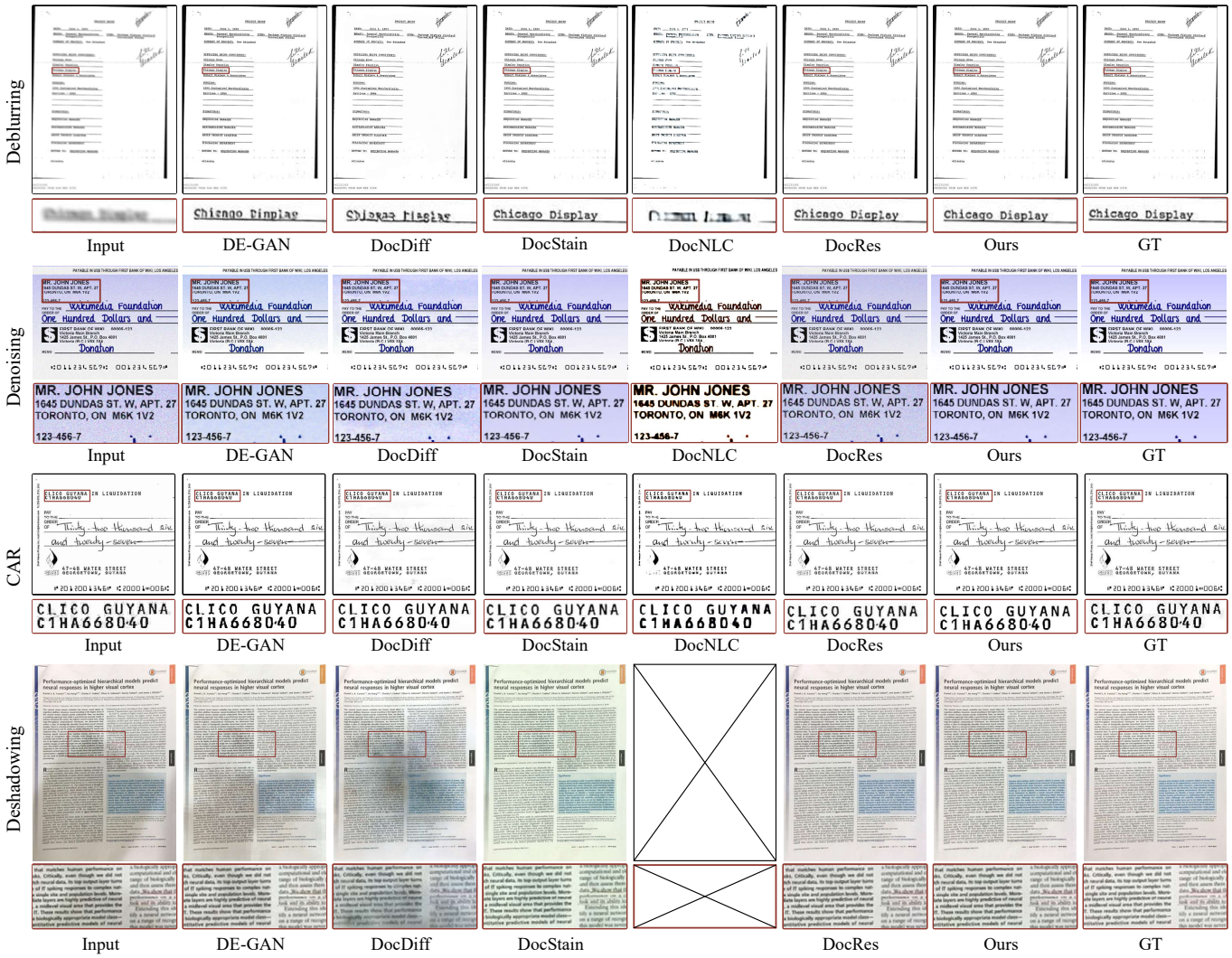

Fig. 14. More qualitative comparison across four restoration tasks: deblurring, denoising, compression artifact reduction (CAR), and deshadowing. The results are reported on the FUNSD$_{Blur}$, BCSD$_{Noise}$, BCSD$_{CAR}$, and Jung's dataset [47], respectively. Note that these datasets feature document styles distinct from the training set. Benefiting from the structure predicted by DASAE and the frequency-domain interaction facilitated by SWIM, our method achieves higher text structural fidelity and better background consistency.

[8] Luwei Tu, Jiawei Wu, Chenxi Wang, Deyu Meng, and Zhi Jin. Fourier-based decoupling network for joint low-light image enhancement and deblurring. *IEEE Transactions on Image Processing*, 2025. 1

[9] Luwei Tu, Jiawei Wu, Xing Luo, and Zhi Jin. Unifying heterogeneous degradations: Uncertainty-aware diffusion bridge model for all-in-one image restoration. *arXiv preprint arXiv:2601.21592*, 2026. 1

[10] Wanglong Lu, Jikai Wang, Tao Wang, Kaihao Zhang, Xianta Jiang, and Hanli Zhao. Visual style prompt learning using diffusion models for blind face restoration. *Pattern Recognition*, 161:111312, 2025. 1

[11] Zinuo Li, Xuhang Chen, Chi-Man Pun, and Xiaodong Cun. High-resolution document shadow removal via a large-scale real-world dataset and a frequency-aware shadow erasing net. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12415–12424, 2023. 1, 3

[12] Zongyuan Yang, Baolin Liu, Yongping Xiong, Lan Yi, Guibin Wu, Xiaojun Tang, Ziqi Liu, Junjie Zhou, and Xing Zhang. DocDiff: Document enhancement via residual diffusion models. In *ACM International Conference on Multimedia (MM)*, pages 2795–2806, 2023. 1, 3, 8, 10, 11, 16

[13] Jiaxin Zhang, Dezhi Peng, Chongyu Liu, Peirong Zhang, and Lianwen Jin. DocRes: A generalist model toward unifying document image restoration tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15654–15664, 2024. 1, 4, 7, 8, 10, 11, 16

[14] Ruilu Wang, Yang Xue, and Lianwen Jin. DocNLC: A document image enhancement framework with normalized and latent contrastive representation for multiple degradations. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pages 5563–5571, 2024. 1, 3, 8, 10, 11, 16

[15] Stéphane Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989. 2, 6, 13

[16] Jie Xiao, Xueyang Fu, Man Zhou, Hongjian Liu, and Zheng-Jun Zha. Random shuffle transformer for image restoration. In *International Conference on Machine Learning (ICML)*, 2023. 2

[17] Jie Xiao, Xueyang Fu, Yurui Zhu, and Zheng-Jun Zha. Bayesian window transformer for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):2431–2444, 2026. 2

[18] Jie Xiao, Xueyang Fu, Yurui Zhu, Dong Li, Jie Huang, Kai Zhu, and Zheng-Jun Zha. Homoformer: Homogenized transformer for image shadow removal. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25617–25626, 2024. 2

[19] Hala Neji, Mohamed Ben Halima, Tarek M. Hamdani, Javier Nogueras-Iso, and Adel M. Alimi. Blur2Sharp: A GAN-based model for document image deblurring. *International Journal of Computational Intelligence Systems*, 14(1):1315–1321, 2021. 2

[20] Sarayut Gonwirat and Olarik Surinta. DeblurGAN-CNN: Effective image denoising and recognition for noisy handwritten characters. *IEEE*

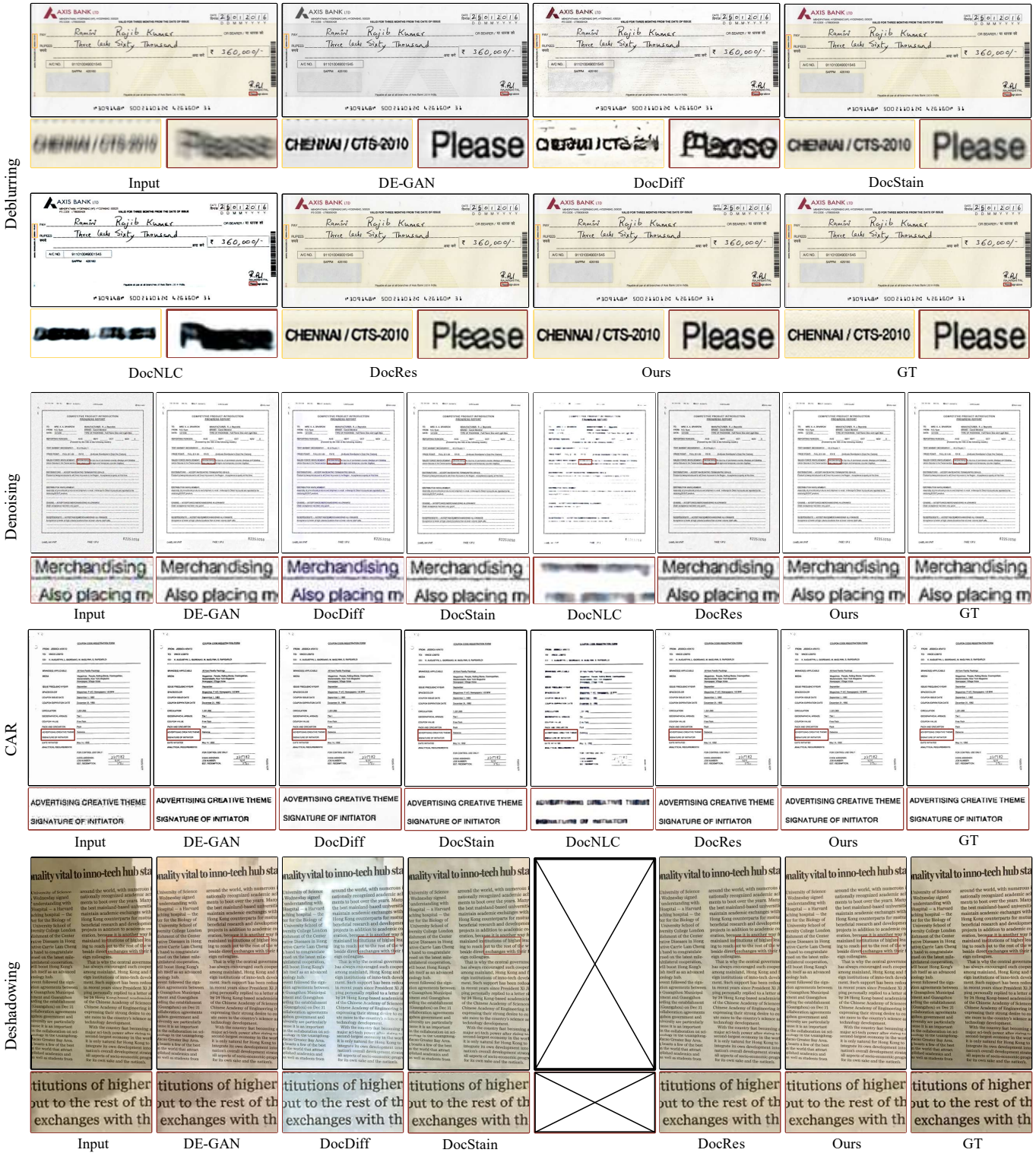

Fig. 15. Qualitative comparison of generalization capabilities across four image restoration tasks: deblurring, denoising, compression artifact reduction (CAR), and deshadowing, evaluated on the BCSD$_{blur}$, FUNSD$_{noise}$, FUNSD$_{CAR}$, and OSR datasets, respectively. The evaluation is conducted on document images with styles distinct from the training set. Leveraging our degradation-aware structure-guided wavelet modulation, our DocPure exhibits enhanced generalization to text patterns and complex backgrounds.

Access, 10:90133–90148, 2022. 2

[21] Jianhan Mei, Ziming Wu, Xiang Chen, Yu Qiao, Henghui Ding, and Xudong Jiang. DeepDeblur: Text image recovery from blur to sharp. *Multimedia Tools and Applications*, 78(13):18869–18885, 2019. 2

[22] Kai Zhang, Wangmeng Zuo, and Lei Zhang. FFDNet: Toward a fast and flexible solution for CNN-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018. 3

[23] Hala Neji, Mohamed Ben Halima, Javier Nogueras-Iso, Tarek M. Hamdani, Javier Lacasta, Habib Chabchoub, and Adel M. Alimi. Doc-Attentive-GAN: Attentive GAN for historical document denoising. *Mul-*

timedia Tools and Applications, 83(18):55509–55525, 2024. 3

[24] Xiaoshuai Zhang, Wenhan Yang, Yueyu Hu, and Jiaying Liu. DMCNN: Dual-domain multi-scale convolutional neural network for compression artifacts removal. In *IEEE International Conference on Image Processing (ICIP)*, pages 390–394, 2018. 3

[25] Jianqi Ma, Shi Guo, and Lei Zhang. Text prior guided scene text image super-resolution. *IEEE Transactions on Image Processing*, 32:1341–1353, 2023. 3

[26] Shihao Zhao, Wanglong Lu, Binhao Wang, Tao Wang, Kaihao Zhang, and Hanli Zhao. Uhdres: Ultra-high-definition image restoration via dual-domain decoupled spectral modulation. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2026. 3

[27] Jiawei Liu, Qiang Wang, Huijie Fan, Jiandong Tian, and Yandong Tang. A shadow imaging bilinear model and three-branch residual network for shadow removal. *IEEE Transactions on Neural Networks and Learning Systems*, 35(11):15857–15871, 2024. 3

[28] Fan Yang, Nanfeng Jiang, Da-Han Wang, Xu-Yao Zhang, Yun Wu, and Shunzhi Zhu. ADR-Net: Attention-oriented detail recovery network for document image shadow removal. *Knowledge-Based Systems*, 329:114228, 2025. 3

[29] Ling Zhang, Yinghao He, Qing Zhang, Zheng Liu, Xiaolong Zhang, and Chunxia Xiao. Document image shadow removal guided by color-aware background. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1818–1827, 2023. 3, 7, 8, 14

[30] Yuhi Matsuo, Naofumi Akimoto, and Yoshimitsu Aoki. Document shadow removal with foreground detection learning from fully synthetic images. In *IEEE International Conference on Image Processing (ICIP)*, pages 1656–1660, 2022. 3, 7, 8

[31] Mohamed Ali Souibgui and Yousri Kessentini. DE-GAN: A conditional generative adversarial network for document enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1180–1191, 2022. 3, 8, 10, 11, 16

[32] Wanglong Lu, Lingming Su, Jingjing Zheng, Vinícius Veloso de Melo, Farzaneh Shoeleh, John Hawkin, Terrence Tricco, Hanli Zhao, and Xianta Jiang. TextDoctor: Unified document image inpainting via patch pyramid diffusion models. *arXiv preprint arXiv:2503.04021*, 2025. 3

[33] Mingxian Li, Hao Sun, Yingtie Lei, Xiaofeng Zhang, Yihang Dong, Yilin Zhou, Zimeng Li, and Xuhang Chen. High-fidelity document stain removal via a large-scale real-world dataset and a memory-augmented transformer. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7614–7624, 2025. 3, 8, 10, 11, 16

[34] Fangmin Zhao, Weichao Zeng, Zhenhang Li, Dongbao Yang, Binbin Li, Xiaojun Bi, and Yu Zhou. Uni-DocDiff: A unified document restoration model based on diffusion. In *ACM International Conference on Multimedia (MM)*, pages 8204–8213, 2025. 4

[35] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5718–5729, 2022. 4, 7

[36] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018. 4

[37] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Rethinking atrous convolution for semantic image segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5615–5623, 2017. 5

[38] Zewen Chi, Li Dong, Shaohan Huang, Damai Dai, Shuming Ma, Barun Patra, Saksham Singhal, Payal Bajaj, Xia Song, et al. On the representation collapse of sparse mixture of experts. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 34600–34613, 2022. 5

[39] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *International Workshop on Deep Learning in Medical Image Analysis*, pages 240–248, 2017. 6

[40] Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. In *International Conference on Medical Imaging with Deep Learning (MIDL)*, pages 285–296, 2019. 6

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5998–6008, 2017. 6

[42] Yuning Cui, Syed Waqas Zamir, Salman Khan, Alois Knoll, Mubarak Shah, and Fahad Khan. AdaIR: Adaptive all-in-one image restoration via frequency mining and modulation. In *International Conference on Learning Representations (ICLR)*, pages 101306–101327, 2025. 6, 7

[43] Michal Hradiš, Jan Kotera, Pavel Zemcık, and Filip Šroubek. Convolutional neural networks for direct text deblurring. In *British Machine Vision Conference (BMVC)*, volume 10, pages 1–13, 2015. 7

[44] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. FUNSD: A dataset for form understanding in noisy scanned documents. In *International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6, 2019. 7

[45] Muhammad Saif Ullah Khan. A novel segmentation dataset for signatures on bank checks. *arXiv preprint arXiv:2104.12203*, 2021. 7

[46] Bingshu Wang and C. L. Philip Chen. Local water-filling algorithm for shadow detection and removal of document images. *Sensors*, 20(23):6929, 2020. 7, 8

[47] Seungjun Jung, Muhammad Abul Hasan, and Changick Kim. Water-filling: An efficient algorithm for digitized document shadow removal. In *Asian Conference on Computer Vision (ACCV)*, pages 398–414, 2018. 7, 8, 14, 17

[48] Jaakko Sauvola and Matti Pietikäinen. Adaptive document image binarization. *Pattern Recognition*, 33(2):225–236, 2000. 8

[49] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. 8

[50] I. Loshchilov and F. Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017. 8

[51] Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, et al. PaddleOCR-VL: Boosting multilingual document parsing via a 0.9b ultra-compact vision-language model. *arXiv preprint arXiv:2510.14528*, 2025. 9

[52] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*, 2023. 9

[53] Ingrid Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 41:909–996, 1988. 13

[54] Hanli Zhao, Binhao Wang, Shihao Zhao, Tao Wang, Kaihao Zhang, and Wanglong Lu. Echosr: Efficient context harnessing for lightweight image super-resolution. *Information Fusion*, 135:104471, 2026. 14