DocRes: A Generalist Model Toward Unifying Document Image Restoration Tasks

Jiaxin Zhang[1,2], Dezhi Peng[1], Chongyu Liu[1], Peirong Zhang[1], Lianwen Jin[1,2,*]
[1] South China University of Technology
[2] INTSIG-SCUT Joint Lab on Document Analysis and Recognition

Abstract

Document image restoration is a crucial aspect of Document AI systems, as the quality of document images significantly influences the overall performance. Prevailing methods address distinct restoration tasks independently, leading to intricate systems and the incapability to harness the potential synergies of multi-task learning. To overcome this challenge, we propose DocRes, a generalist model that unifies five document image restoration tasks including dewarping, deshadowing, appearance enhancement, deblurring, and binarization. To instruct DocRes to perform various restoration tasks, we propose a novel visual prompt approach called ***D****ynamic* ***T****ask-****S****pecific* ***Prompt*** *(DTSPrompt). The DTSPrompt for different tasks comprises distinct prior features, which are additional characteristics extracted from the input image. Beyond its role as a cue for task-specific execution, DTSPrompt can also serve as supplementary information to enhance the model's performance. Moreover, DTSPrompt is more flexible than prior visual prompt approaches as it can be seamlessly applied and adapted to inputs with high and variable resolutions. Experimental results demonstrate that DocRes achieves competitive or superior performance compared to existing state-of-the-art task-specific models. This underscores the potential of DocRes across a broader spectrum of document image restoration tasks. The source code is publicly available at* https://github.com/ZZZHANG-jx/DocRes.

1. Introduction

Photographed or scanned document images frequently manifest various forms of degradation, encompassing geometric distortion, shadows, bleed-through, stains, and more. These degradations pose substantial challenges for existing document analysis and recognition systems, and also significantly compromise the visual appeal and legibility of documents, as underscored in recent studies [5, 9, 23, 59, 70].

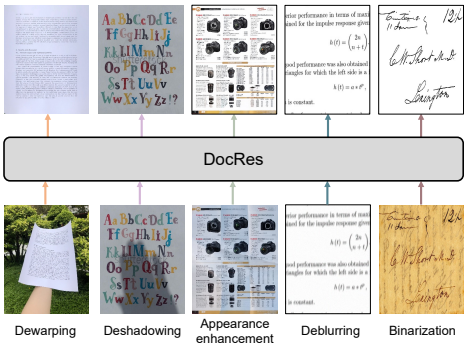

Figure 1. DocRes is a generalist model that unifies five document image restoration tasks, including tasks of dewarping, deshadowing, appearance enhancement, deblurring, and binarization.

The restoration of document images, addressing objectives like flattening documents, shadow removal, clean appearance restoration, deblurring, or text segmentation, holds both academic and practical significance. Existing approaches [5, 9, 33, 59, 62] typically treat these restoration tasks separately, relying on models specifically designed for each task. While effective in achieving commendable performance, this paradigm results in a system that requires the design of multiple models and extensive maintenance. Moreover, it fails to leverage the potential synergies of multi-task learning.

To tackle this problem, motivated by recent pioneering works [3, 24, 30, 38, 57, 58] that unify various vision tasks, we seek to explore a generalist model for document image restoration. As illustrated in Fig. 1, our proposed DocRes seeks to unify five document image restoration tasks: dewarping, deshadowing, appearance enhancement, deblurring, and binarization. To empower such a generalist model to perform specific tasks and generate desired outputs, it is often crucial to convey instructional information to the model. Existing visual generalist models accomplish this by drawing inspiration from advancements in Natural Language Processing (NLP) and leveraging prompt learning

*Corresponding author

techniques. Notably, studies like [3, 52, 57] transform the vision task into an NLP one by discretizing continuous vision output and using discrete tokens as prompts. However, the autoregressive decoding paradigm they employ is inherently less effective for low-level tasks.

In the realm of unifying image-to-image tasks, recent methods [1, 29, 32, 58] have suggested vision-centric prompts, known as visual prompts, which exhibit promising potential across various vision tasks, including image restoration. Among them, approaches like [1, 29, 58] propose using a pair of input/output samples as a visual prompt to guide the model to perform the corresponding task. Nevertheless, these methods exhibit reduced efficiency as they necessitate an additional pair of samples during inference, making them less suitable for low-level tasks involving high-resolution images. ProRes [32] introduces a more straightforward visual prompt method, wherein a matrix composed of learnable parameters, matching the input's shape, is assigned for each task and added to the input as a visual prompt. However, the training process of ProRes is intricate, which requires initializing each visual prompt by pre-training on task-specific models. Additionally, some of the above visual prompt methods rely on Mask Image Modeling (MIM) and require the ViT [8] framework. However, ViT typically demands that input images during testing and training have the same resolution, and due to memory constraints, ViT struggles with high-resolution images. This makes it challenging for these methods to adapt to low-level tasks, which typically involve patch training, whole-image testing, as well as variable and high-resolution inputs.

Recognizing the limitations of the aforementioned prompt methods, we introduce a new visual prompt for DocRes called the **D**ynamic **T**ask-**S**pecific **Prompt** (DTSPrompt). The inspiration behind DTSPrompt is rooted in prevalent practices observed in existing document image restoration endeavors, where certain prior features extracted from the input image are typically employed to enhance model's performance, such as background images for shadow removal [26, 71] and text content masks for dewarping [11, 17, 68]. Specifically, the DTSPrompt for different tasks comprises distinct prior features based on the characteristics of each task, where prior features we adopted include document segmentation masks, binarization results, gradient maps, and so on. DTSPrompt can not only serve as an effective cue for the model to determine which task to perform but also function as supplementary information to enhance the performance for the corresponding task. DTSPrompt can be seamlessly applied to various existing restoration networks, rather than limited to ViT. This enables DocRes to handle high and variable resolutions encountered in document restoration.

Experiments demonstrate that, without additional complex network designs, DocRes, as a generalist model, achieves competitive or even superior performance on various benchmarks compared to existing well-established and carefully designed task-specific methods.

In summary, our contributions are as follows:

- To the best of our knowledge, we are the first to explore generalist models for unifying restoration of document images, which serves as a pioneering effort in this field.
- We propose a simple yet highly effective visual prompt approach termed DTSPrompt, which extracts different prior features from the input image to create prompts. It can guide the model to distinguish different tasks, provide supplementary information to improve performance and accommodate both high and variable resolutions.
- Empowered by DTSPrompt, our DocRes achieves competitive or even better performance compared to task-specific methods across various benchmarks.

2. Related works

2.1. Document Image Restoration

Dewarping [5, 10, 11, 33, 68] addresses the elimination of geometric distortions such as curves and crumples, which not only hinder OCR engine performance [4, 56, 70] but also degrade document readability. Deshadowing [25, 26, 71] focuses on removing shadows, a common occurrence in photographed document images, to produce shadow-free documents. Appearance enhancement, also known as illumination correction [6, 59, 70], goes beyond specific appearance degradations, striving to restore a clean appearance akin to digital-born PDFs. This is valuable as it significantly enhances document readability and aesthetics. Deblurring [14, 50, 61] aims to eliminate blurriness and restore a clear image. Binarization [62] involves segmenting foreground text from document images, a critical task for applications primarily focused on text content, usually obscured by stains, artifacts, black margins, or weak contrast.

Existing approaches [5, 26, 59, 62, 68, 70, 71] typically treat these restoration tasks independently, resulting in a complex document image restoration system and the inability to harness the potential synergies among tasks. While recent efforts [50, 51, 61] aim to tackle several tasks with a unified network architecture, they still require individual training and separate models for each task.

2.2. All-in-one image restoration

Unlike the domain of document image restoration that lacks unification, recent studies [24, 29, 32, 54, 58, 67] have made pioneering strides in developing generalist models for natural scene image restoration. They encompass tasks such as weather effect removal, low-light image enhancement, denoising, and deblurring. Existing multi-task generalist models can be broadly classified into two categories based on

whether the task is explicitly specified for the model during inference: task-agnostic and task-oriented.

Task-agnostic. Task-agnostic methods [21, 22, 39, 54, 67] do not require users to specify the task type, but they are less flexible and cannot handle a broader range of tasks, typically being confined to specific domains like weather effect removal. This arises because some tasks share similar inputs but demand distinct outputs, leading to ambiguity in the learning process. The task-agnostic setup proves inappropriate for unifying document image restoration since tasks like dewarping, deshadowing, and appearance enhancement share similar inputs but demand distinct outputs.

Task-oriented. To achieve a task-oriented generalist model, explicit task information needs to be introduced. Some approaches [3, 24, 30, 57, 58] discretize the continuous vision output, use discrete tokens as task prompts, and use the Transformer decoder for autoregressive prediction. These methods are more suited for visual understanding tasks, such as detection, captioning, and visual question answering but are inappropriate for low-level tasks involving high-resolution outputs. More recently, there have been methods proposing vision-centric prompts [1, 29, 32, 58] for unification purposes, known as visual prompts, showing promising potential in various vision tasks, including image restoration. Among them, [1, 29, 58] use an input/output sample of a specific task as a prompt, resembling the paradigm used in inpainting tasks where surrounding pixel information is learned to fill in missing pixel positions. Nevertheless, the additional input leads to inefficiency and limitation of image resolution. ProRes [32] employs learnable parameters with the same shape of the input image as a visual prompt and adds it pixel-wise to the input. While effectively guiding the model, the prompt for each task needs to be trained on task-specific models for initialization, resulting in a complex training pipeline.

Moreover, all these visual prompt methods mentioned above are limited to the ViT [8] framework, making them unable to adapt to the variable resolutions in restoration tasks. The quadratic computational complexity also restricts the input resolution. For example, the input resolution for ProRes [32] and Painter [58] is limited to 448×448, which is insufficient for document images with resolutions commonly exceeding 1K. While block-wise processing could be employed, many document image restoration tasks heavily depend on global information, such as deshadowing and binarization, making this strategy unfeasible. In contrast, our DTSPrompt method is not confined to the ViT framework and can be applied to various more flexible restoration networks to form our DocRes model.

3. Methodology

As depicted in Fig. 2, the input document image undergoes an initial processing step by the DTSPrompt generator to generate task-specific DTSPrompts, which are composed of various prior features extracted from the input image. Such DTSPrompts are instrumental in guiding the restoration network to execute distinct tasks while simultaneously enhancing overall performance. In the following sections, we first provide a detailed exploration of the process involved in obtaining various prior features to construct the DTSPrompt for each task. Then we introduce our prompt fusion approach and the selection of the restoration network.

3.1. Dynamic task-specific prompt

Prevailing visual generalist models [29, 32, 52, 57, 58] determine prompts solely based on the task, and are independent of the input image:

$$Prompt = f(\text{ task }). \tag{1}$$

In contrast, our DTSPrompt dynamically adapts to the input image $I_s \in \mathbb{R}^{h \times w \times 3}$:

$$DTSPrompt = G\left(I_s, \text{ task }\right) \in \mathbb{R}^{h \times w \times 3}. \tag{2}$$

Here, G represents our DTSPrompt generator (depicted in Fig. 2), which is responsible for extracting prior features from I_s based on the specified task. In the following, we provide a detailed explanation of how the DTSPrompt for each task is constructed. Visual results of DTSPrompts for each task are presented in Fig. 3.

Dewarping. Existing document dewarping methods [11, 17, 23] often use text line masks or text block masks to assist the dewarping model, making it more attentive to the dewarping of regions with meaningful content. Document mask [9, 11, 23, 68] is commonly employed to enhance the model's understanding of document boundaries and reduce the learning difficulty by decoupling the margin removal and content rectification processes. Here, we choose the simplest document mask as our prior feature for the dewarping task. This mask, denoted as $P_m(I_s) \in \mathbb{R}^{h \times w}$, is obtained by directly using an existing document segmentation model proposed in [68] (noted that any other existing document segmentation models [9, 11] can also be employed).

Furthermore, considering that predicting the backward map in dewarping task is inherently a problem related to coordinate positions, inspired by [5, 27], we introduce the x-coordinate and y-coordinate as additional prior features, denoted as $P_{cx} \in \mathbb{R}^{h \times w}$ and $P_{cy} \in \mathbb{R}^{h \times w}$, respectively. They represent the coordinate values of the pixel at (i, j), i.e., $P_{cx}(i, j) = i$ and $P_{cy}(i, j) = j$, which enable better perception of positional information. The DTSPrompt for dewarping task is obtained by concatenating these prior features along the channel dimension:

$$G\left(I_s, \text{“}dewarp\text{”}\right) = \left[P_m\left(I_s\right), P_{cx}, P_{cy}\right]. \tag{3}$$

Deshadowing. For the deshadowing task, the document background is commonly considered as a prior feature to enhance performance [26, 28, 71]. Here, we adopt

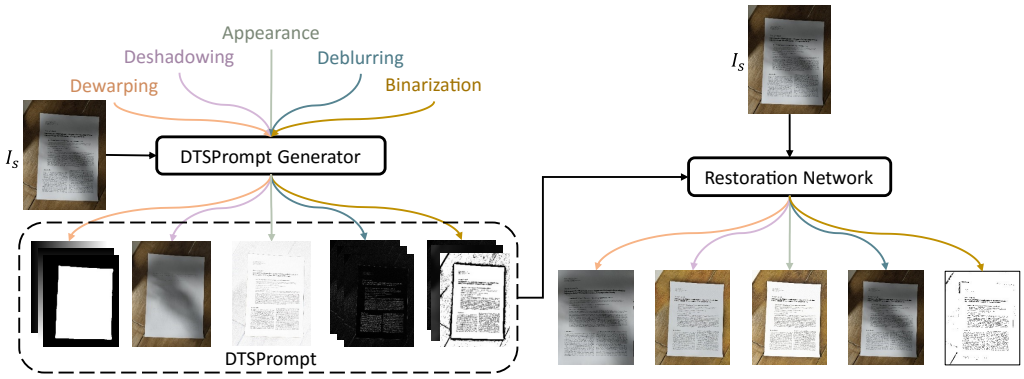

Figure 2. The overall pipeline for DocRes. The document image to be restored, denoted as I_s, is initially fed into the DTSPrompt generator, which extracts specific prior features based on the task to form the DTSPrompt. Alongside I_s, DTSPrompt is input into the restoration network. It serves not only as a guidance for the restoration network on the particular task to be performed but also functions as auxiliary information derived from I_s to improve performance.

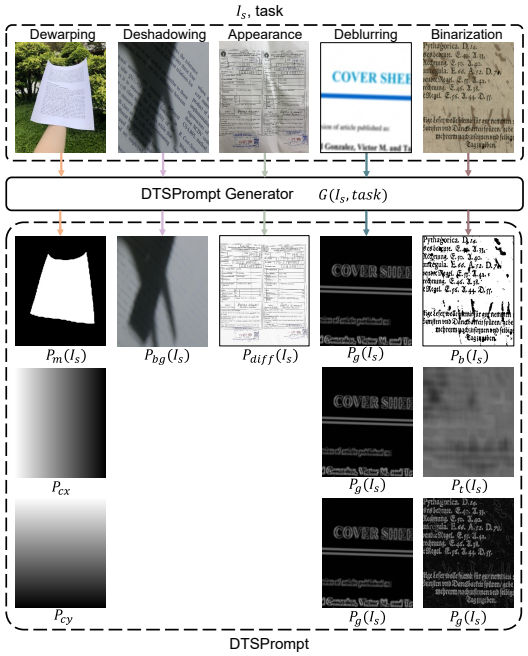

Figure 3. The DTSPrompt for different tasks is composed of distinct prior features. Most of these prior features are extracted from the input image, making them dynamic. Zoom in for the best view.

the document background with shadows as our prior feature. To obtain this background from the document image, we initially employ dilation operations to eliminate the textual content within the document. Subsequently, we use a median filter to smooth out artifacts introduced due to incomplete removal. We represent this whole process as $P_{bg}(I_s) \in \mathbb{R}^{h\times w\times 3}$. The DTSPrompt for the deshadowing task is thus formulated as

$$G\left(I_s, \text{“}deshadow\text{”}\right) = P_{bg}\left(I_s\right). \tag{4}$$

Appearance enhancement. Current methods [6, 59, 70] commonly adopt background light, shadow map, or white-balance kernel as prior features to facilitate appearance enhancement based on the concept of intrinsic images. However, accurately obtaining these prior features is challenging and usually requires an extra model for training and predicting. For simplicity, we leverage the discrepancy between the original image and the document background $P_{bg}(I_s)$ as our prior feature for this task, which can serve as initial enhancement guidance for the model:

$$P_{diff}(I_s) = 255 - abs(I_s - P_{bg}(I_s)), \tag{5}$$

$$G\left(I_s, \text{“}appearance\text{”}\right) = P_{diff}\left(I_s\right). \tag{6}$$

Deblurring. The gradient distribution is a prior feature widely used in traditional optimization-based deblurring methods [20], which is typically employed as regularization information to constrain the solution space of the optimization function. Here, we use the gradient map of the input image $P_g\left(I_s\right) \in \mathbb{R}^{h\times w}$ as an additional input, aiming for the model to implicitly learn gradient prior information, rather than utilizing a gradient distribution prior to constrain the solution space of the output. The DTSPrompt for the deblurring task is expressed as

$$G\left(I_s, \text{“}deblur\text{”}\right) = \left[P_g\left(I_s\right), P_g\left(I_s\right), P_g\left(I_s\right)\right]. \tag{7}$$

Binarization. Integrating prior features as supplementary for performance enhancement is widely adopted in binarization task [16, 53, 62], and the effectiveness has been extensively demonstrated. For this task, we employ the Sauvola binarization algorithm [49] to yield the initial binarization outcome and threshold map as our prior features, which is denoted as $P_b(I_s) \in \mathbb{R}^{h\times w}$ and $P_t(I_s) \in \mathbb{R}^{h\times w}$,

respectively. Furthermore, we also incorporate gradient information as an additional prior feature for this task. The DTSPrompt for binarization task is formulated as

$$G\left(I_s, \text{“}binarize\text{”}\right) = \left[P_b\left(I_s\right), P_t\left(I_s\right), P_g\left(I_s\right)\right]. \quad (8)$$

3.2. Prompt fusion and restoration network

Exploring how to integrate the acquired prompt information into the network holds significant merit. A well-conceived fusion method has the potential to substantially enhance overall performance. However, the primary focus of this paper lies in evaluating the efficacy and potentials associated with DTSPrompt, rather than delving into the intricacies of designing a complex network structure. In line with this objective, we adopt a straightforward fusion approach to seamlessly incorporate DTSPrompt into the restoration network. Specifically, we opt for concatenating DTSPrompt and I_s along the channel dimension to construct a new input $\in \mathbb{R}^{h\times w\times 6}$ for the restoration network.

Due to the simplicity of DTSPrompt, we have the flexibility to choose from various restoration networks. In this case, we opt for the off-the-shelf Restormer [65] without modifications to form our DocRes. With such a restoration network, DocRes can support inputs of up to 1600×1600 and adapt to inputs with variable resolutions. It's noteworthy that other restoration networks can be employed interchangeably since DTSPrompt does not require specific modules within the network.

4. Experiments

4.1. Datasets

Dewarping. We adopt the Doc3D [5] dataset and the DIR300 benchmark [11] for the training and testing, respectively. Doc3D is a synthetic dataset comprising 100K samples, which includes geometrically distorted document images and corresponding backward maps. DIR300 is a real-world benchmark with 300 geometrically distorted images and corresponding flat ground-truths.

Deshadowing. The training set for this task consists of 14,200 synthetic images from FSDSRD [35] and 4,371 real images from the training set of RDD [71]. We use Jung's dataset [18] (87 images), Kligler's dataset [19] (300 images) and OSR [55] (237 images) to form our testing set.

Appearance enhancement. The training set for this task contains 90K synthetic images from the Doc3DShade [6] dataset and 450 real-world images from the RealDAE [70] training set. 150 images from the testing set of RealDAE and 130 images from DocUNet [33] are used for evaluation. As introduced in [5, 9], the degraded images in DocUNet should be aligned to the flat ground truths before the evaluation of this task. Following [70], we achieve alignment by using the document alignment model [69] rather than some dewarping models [9, 33, 68], which can result in better alignment and thus provide a more accurate evaluation. We denote the aligned dataset as DocUNet*.

Deblurring. The Text Deblur Dataset (TDD) [14] consists of 66K training samples, from which we randomly select 40K samples to train our model. The 1.6K testing samples from TDD form the testing set of this task.

Binarization. We use DIBCO'18 [46] as our testing set. Following [61, 62], the remaining years of (H)-DIBCO datasets [12, 37, 40–45, 47] are used as the training data, and images from Noisy Office dataset [66], Synchromedia Multispectral dataset [13], Persian Heritage Image Binarization dataset [36] and Bickley Diary dataset [7] are also used for training.

4.2. Evaluation metrics

Deshadowing, appearance enhancement, and deblurring tasks adopt the commonly used PSNR and SSIM as evaluation metrics. The evaluation of dewarping incorporates multi-scale structural similarity (MS-SSIM) [60], local distortion (LD) [64] and align distortion (AD) [34]. MS-SSIM builds upon the traditional SSIM by considering multiple scales. LD evaluates dewarping performance by utilizing the offset between the dewarped result and the flat ground truth. AD, an enhancement of LD, refines the evaluation by excluding offset noise in low-textured regions and mitigating the impact of global transformations. the parameters for MS-SSIM, LD, and AD align with those established in prior works [5, 11, 23, 33]. For the binarization task, we employ PSNR, F-measure (FM), and pseudo F-measure (pFM) as our evaluation metrics.

4.3. Implementation details

We train our model on 8 NVIDIA A6000 GPUs for 100,000 steps with a global batch size of 80. AdamW with a weight decay of 5×10^{-4} is adopted. We use the cosine learning rate scheduler with 2×10^{-4} as the maximum learning rate.

Before commencing such unified training, the model undergoes a pre-training phase exclusively on the dewarping task for 50,000 steps to initialize the model. This is for a more stable training purpose, as dewraping significantly differs from other tasks: while it involves coordinates regression, other tasks entail regression of image content.

During the unified training process, the sampling weight for dewarping, deshadowing, appearance enhancement, deblurring, and binarization are all set to 0.2. Apart from the binarization task, which employs the standard cross-entropy loss for its output, all other tasks are supervised using the L1 loss. Images of deshadowing, appearance enhancement, deblurring, and binarization tasks are randomly cropped as patches with a size of 256×256, while images of dewarping task are resized to 256×256 during training.

Table 1. Quantitative comparison results between our all-in-one generalist DocRes model and existing task-specific state-of-the-art models on 5 tasks. From top to bottom, the tasks include dewarping, deshadowing, appearance enhancement, deblurring, and binarization. **Best** results are shown in bold.

Metrics	Datasets	DocGeo [11] *ECCV'22*	Li et al. [23] *ICCV'23*	BGSNet [71] *CVPR'23*	DocShadow [25] *ICCV'23*	UDoc-GAN [59] *MM'22*	GCDRNet [70] *TAI'23*	GDB [62] *PR'24*	DE-GAN [50] *TPAMI'20*	DocDiff [61] *MM'23*	DocRes (**ours**)
MSSSIM↑	DIR300 [11]	**0.6380**	0.6070								0.6264
AD↓		0.242	0.244	-	-	-	-	-	-	-	**0.241**
LD↓		**6.40**	7.68								6.83
SSIM↑	Kligler et al. [19]	-	-	**0.9480**	0.9088	-	-	-	-	-	0.9005
PSNR↑				**29.17**	25.12						27.14
SSIM↑	Jung et al. [18]	-	-	0.9040	0.9005	-	-	-	-	-	**0.9089**
PSNR↑				17.34	21.05						**23.02**
SSIM↑	OSR [55]	-	-	**0.9388**	0.9023	-	-	-	-	-	0.9370
PSNR↑				**22.64**	18.25						21.64
SSIM↑	DocUNet* [33]	-	-	-	-	0.6833	**0.7658**	-	-	-	0.7598
PSNR↑						14.29	17.09				**17.60**
SSIM↑	RealDAE [70]	-	-	-	-	0.7558	**0.9423**	-	-	-	0.9219
PSNR↑						16.43	24.42				**24.65**
SSIM↑	TDD [14]	-	-	-	-	-	-	-	0.9226	0.9559	**0.9723**
PSNR↑									22.24	24.00	**27.35**
FM↑	DIBCO'18 [46]	-	-	-	-	-	-	**91.09**	77.59	88.11	89.82
pFM↑								**94.57**	85.74	90.43	94.33
PSNR↑								**19.92**	16.16	17.92	19.35

Table 2. Quantitative results of the ablation experiments. **Best** results are shown in bold.

Metrics	Datasets	Task specific	Unified		
		Baseline	Baseline	Baseline +Fix prompt	Baseline +DTSPrompt (DocRes)
MSSSIM↑	DIR300 [11]	0.6185	0.5943	0.6030	**0.6264**
AD↓		0.263	0.333	0.302	**0.241**
LD↓		7.18	8.26	7.77	**6.83**
SSIM↑	Kilgler et al. [19]	0.8879	0.8238	0.8977	**0.9005**
PSNR↑		**27.41**	11.95	27.08	27.14
SSIM↑	Jung et al. [18]	0.9025	0.8573	0.9031	**0.9089**
PSNR↑		**23.18**	18.01	22.42	23.02
SSIM↑	OSR [55]	0.9259	0.8634	0.9285	**0.9370**
PSNR↑		20.12	12.35	19.47	**21.64**
SSIM↑	DocUNet* [33]	0.7621	0.7620	**0.7635**	0.7598
PSNR↑		17.43	17.30	17.28	**17.60**
SSIM↑	RealDAE [70]	0.9204	0.9050	0.9200	**0.9219**
PSNR↑		24.21	21.75	24.32	**24.65**
SSIM↑	TDD [14]	**0.9811**	0.9639	0.9690	0.9723
PSNR↑		**28.95**	25.85	26.64	27.35
FM↑	DIBCO'18 [46]	76.57	74.33	77.15	**89.82**
pFM↑		79.51	76.15	81.18	**94.33**
PSNR↑		14.64	14.47	15.28	**19.35**

4.4. Results

Comparisons with SOTA task-specific models. We conduct a comprehensive comparison between our proposed DocRes and existing meticulously designed task-specific methods. Specifically, for the dewarping task, we benchmark DocRes against the current state-of-the-art method, DocGeo [11], and the recent model by Li et al. [23]. In the deshadowing domain, we compare DocRes with the latest SOTA models, namely BGSNet [71] (utilizing the model trained on the RDD dataset provided by the authors) and DocShadow [25] (based on the model trained on the SD7K dataset provided by the authors). For appearance enhancement, we assess DocRes against UDoc-GAN [59] and GCDRNet [70]. The binarization task involves a comparison with the current SOTA method, GDB [62]. Additionally, we contrast our approach with DocDiff [61] and DE-GAN [50], both of which aim to unify multiple tasks within a single network structure but still necessitate separate training for each task, resulting in the need for multiple models.

Results across multiple benchmark datasets for the five tasks are presented in Table 1. It can be seen that DocRes not only competes with existing task-specific SOTA models but also surpasses them in several instances. DocRes achieves new records in certain metrics for benchmark datasets related to dewarping, deshadowing, deblurring, and appearance enhancement tasks. Even for binarization tasks, where the dedicated SOTA model GDB still holds the top position, DocRes exhibits performance closely trailing behind it. In contrast to existing unified-structure methods like DE-GAN and DocDiff, which still require separate training for each task, DocRes demonstrates significant advantages. Visualized results on these benchmarks from DocRes are showcased in Fig. 4.

While there remains room for improvement compared to some of these well-designed specialized models, it's important to underscore that our primary objective in this paper is not to achieve SOTA performance on every task. Instead, the focus is on evaluating the efficacy and potential of the unified DocRes approach. Further research, for example, can explore more sophisticated prompt fusion mechanisms to achieve SOTA performance in each task.

Ablation studies. In this subsection, we conduct ablation studies to evaluate the effectiveness of our DTSPrompt. Restormer [65] is treated as our baseline model. We first individually train it on each task, obtaining task-specific results shown in the third column of Table 2. As a state-of-the-art image restoration network, Restormer demonstrates proficiency across various tasks when trained in isolation. However, when training it in a unified model setting (the fourth column of Table 2), we see a significant performance decline across almost all benchmarks. This decline is at-

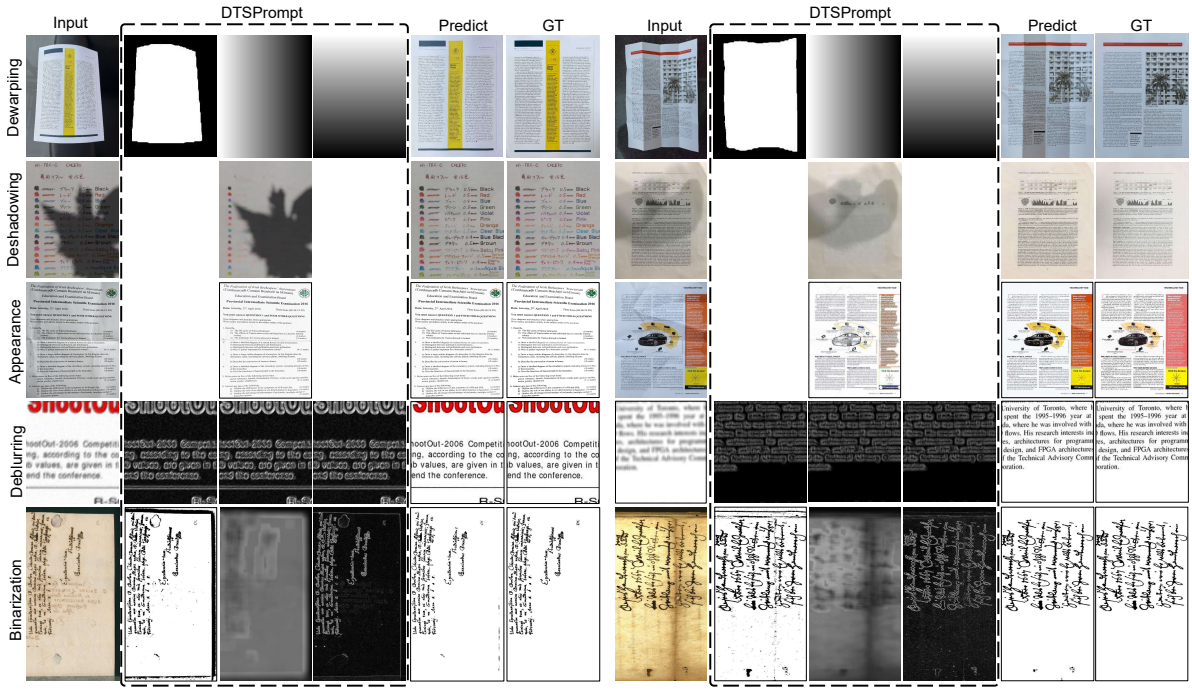

Figure 4. Visualizations showcasing inputs, DTSPrompt, and restoration results from DocRes, and ground truths across various tasks, including dewarping, deshadowing, appearance enhancement, deblurring, and binarization. Zoom in for the best view.

Figure 5. For the same input, we employ different DTSPrompts to validate the controllability of DocRes. The results in the last column indicate that using DocRes enables the complete restoration process for photographed documents, achieving the desired final results for users. Zoom in for the best view.

tributed to the similarity in input across tasks, coupled with distinct output requirements, leading to confusion in the model's learning process.

Additionally, we explore the effectiveness of fixed prompts. A fixed prompt for each task is a $h \times w \times 3$ matrix, with constant values determined solely by the task and remaining constant regardless of the input image. For example, the fixed prompt for the dewarping task is a $h \times w \times 3$ matrix filled with zeros, while for the deshadowing task, it is filled with ones. We adopt the same fusion approach, where the concatenated result of the fixed prompt and the input image are fed into the restoration network.

Such a simple fixed prompt approach achieves competitive performance compared to task-specific settings in appearance enhancement and deshadowing tasks. There is even a noticeable improvement in the binarization task, which could be attributed to multi-task learning helping mitigate the generalization issue caused by the scarcity of training data for the binarization task. However, for tasks like dewarping and deblurring, the fixed prompt method still experiences significant performance declines, highlighting the challenges of creating an excellent generalist model.

Our DTSPrompt excels in dewarping, deshadowing, deblurring, and binarization tasks compared to the fixed prompt approach. Particularly, there are significant improvements in the dewarping and binarization tasks. From an overall perspective across all tasks, DTSPrompt achieves superior results compared to the task-specific setting, by using only a single model without the need for additional training parameters or structural modifications.

Control ability. In this subsection, we explore the control ability of DTSPrompt. Our focus is on observing how well DocRes can perform the correct restoration task when different DTSPrompts are employed for the same input image. Specifically, we consider three tasks related to the photographed scene: dewarping, deshadowing, and appearance enhancement. The visualization results in the first four

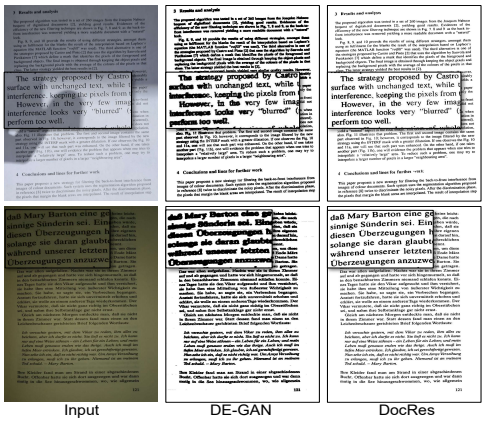

Figure 6. Visualization results when applying models (DE-GAN [50] and our DocRes) to perform binarization on photographed documents, which were merely trained on scanned ancient document binarization data. Zoom in for the best view.

columns of Fig. 5 show that DocRes can accurately perform the corresponding tasks. In addition, we also demonstrate in the last column that a single DocRes model can complete the entire enhancement process of photographed document images and obtain the final results desired by the user.

Generalization. An essential trait of a generalist model lies in its ability to harness synergies among multi-task data to improve overall generalization. In this context, we delve into the generalization capability of DocRes through visualizations. As outlined in Section 4.1, the training data for the binarization task in DocRes primarily consists of scanned documents, particularly ancient ones, presenting a considerable gap from the photographed modern documents [2]. Here, we aim to apply DocRes to binarize photographed documents, introducing unseen noises like blurriness, shadows, low-light conditions, and reflections that are not present in the binarization training set. As shown in Fig. 6, DocRes consistently demonstrates excellent performance on such out-of-domain data. In contrast, DE-GAN's [50] performance significantly deteriorates in the presence of shadow and low-light interference.

Moreover, we extend the evaluation to the deblurring task for photographed document images. Notably, the training data for the deblurring task in DocRes consists exclusively of clean document images. As illustrated in Fig. 7, DocRes exhibits superior deblurring performance on photographed document images compared to DE-GAN [50] and DocDiff [61], both of which were also trained exclusively on the TDD dataset for the deblurring task.

We attribute DocRes's robust out-of-domain generalization capability to its learning of patterns associated with photographed noise through tasks like dewarping, deshadowing, and appearance enhancement.

Figure 7. Visualization results when applying models (DE-GAN [50], DocDiff [61] and our DocRes) to perform the deblurring task on photographed documents, which were merely trained on clean document blurring data. Zoom in for the best view.

5. Discussions and conclusions

This paper presents DocRes, a generalist model designed for unifying document image restoration tasks, including dewarping, deshadowing, appearance enhancement, deblurring, and binarization. The key innovation of DocRes is the incorporation of Dynamic Task-Specific Prompt (DTSPrompt), which leverages prior features to construct visual prompts, acting not only as a guiding cue for specific tasks but also supplying additional information to enhance restoration performance. It can be seamlessly applied to existing restoration networks, resulting in a generalist model that can accommodate input with high or variable resolutions. With the support of DTSPrompt, DocRes achieves performance levels matching or surpassing SOTA task-specific models, without the need for extra training parameters or complex architectural designs. We also illustrate DocRes's controllability in performing different tasks when presented with the same input image and its capacity to generalize to out-of-domain data.

Notably, the DTSPrompt is not confined to the specific tasks explored in this paper. It can be potentially extended to incorporate more diverse prior features [15, 31, 48, 63], such as DCT coefficients, SIFT, JPEG noise, and resampling artifacts, to accommodate a broader range of image restoration tasks. Additionally, it is worth investigating prompt fusion mechanisms for better integrating the DTSPrompt. In summary, this paper successfully attempts to develop a unified multi-task model for document image restoration, inspiring future research of generalist or foundation models for pixel-level image processing tasks.

Acknowledgement

This paper is supported in part by National Natural Science Foundation of China (Grant No.: 61936003), National Natural Science Foundation of China AIGC 2024 Special Project (Grand App. No.: 6244100044), National Key Research and Development Program of China (2022YFC3301703)

References

[1] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *NeurIPS*, 35:25005–25017, 2022. 2, 3

[2] Rodrigo Bernardino, Rafael Dueire Lins, and Ricardo da Silva Barboza. A quality, size and time assessment of the binarization of documents photographed by smartphones. *Journal of Imaging*, 9(2):41, 2023. 8

[3] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey E Hinton. A unified sequence interface for vision tasks. *NeurIPS*, 35:31333–31346, 2022. 1, 2, 3

[4] Hiuyi Cheng, Peirong Zhang, Sihang Wu, Jiaxin Zhang, Qiyuan Zhu, Zecheng Xie, Jing Li, Kai Ding, and Lianwen Jin. M6Doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis. In *CVPR*, pages 15138–15147, 2023. 2

[5] Sagnik Das, Ke Ma, Zhixin Shu, Dimitris Samaras, and Roy Shilkrot. DewarpNet: Single-image document unwarping with stacked 3D and 2D regression networks. In *ICCV*, pages 131–140, 2019. 1, 2, 3, 5

[6] S Das, H Ma Sial, R Baldrich, M Vanrell, and D Samaras. Intrinsic decomposition of document images in-the-wild. In *BMVC*, 2020. 2, 4, 5

[7] Fanbo Deng, Zheng Wu, Zheng Lu, and Michael S Brown. Binarizationshop: a user-assisted software suite for converting old documents to black-and-white. In *JCDL*, pages 255–258, 2010. 5

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 2, 3

[9] Hao Feng, Yuechen Wang, Wengang Zhou, Jiajun Deng, and Houqiang Li. DocTr: Document image transformer for geometric unwarping and illumination correction. In *ACM MM*, pages 273–281, 2021. 1, 3, 5

[10] Hao Feng, Wengang Zhou, Jiajun Deng, Qi Tian, and Houqiang Li. Docscanner: Robust document image rectification with progressive learning. *arXiv preprint arXiv:2110.14968*, 2021. 2

[11] Hao Feng, Wengang Zhou, Jiajun Deng, Yuechen Wang, and Houqiang Li. Geometric representation learning for document image rectification. In *ECCV*, pages 475–492, 2022. 2, 3, 5, 6

[12] Basilis Gatos, Konstantinos Ntirogiannis, and Ioannis Pratikakis. ICDAR 2009 document image binarization contest (DIBCO 2009). In *ICDAR*, pages 1375–1382, 2009. 5

[13] Rachid Hedjam, Hossein Ziaei Nafchi, Reza Farrahi Moghaddam, Margaret Kalacska, and Mohamed Cheriet. ICDAR 2015 contest on multispectral text extraction. In *ICDAR*, pages 1181–1185, 2015. 5

[14] Michal Hradiš, Jan Kotera, Pavel Zemcık, and Filip Šroubek. Convolutional neural networks for direct text deblurring. In *BMVC*, 2015. 2, 5, 6

[15] Fangjun Huang, Jiwu Huang, and Yun Qing Shi. Detecting double JPEG compression with the same quantization matrix. *IEEE TIFS*, 5(4):848–856, 2010. 8

[16] Fuxi Jia, Cunzhao Shi, Kun He, Chunheng Wang, and Baihua Xiao. Degraded document image binarization using structural symmetry of strokes. *Pattern Recognition*, 74: 225–240, 2018. 4

[17] Xiangwei Jiang, Rujiao Long, Nan Xue, Zhibo Yang, Cong Yao, and Gui-Song Xia. Revisiting document image dewarping by grid regularization. In *CVPR*, pages 4543–4552, 2022. 2, 3

[18] Seungjun Jung, Muhammad Abul Hasan, and Changick Kim. Water-filling: An efficient algorithm for digitized document shadow removal. In *ACCV*, pages 398–414, 2018. 5, 6

[19] Netanel Kligler, Sagi Katz, and Ayellet Tal. Document enhancement using visibility detection. In *CVPR*, pages 2374–2382, 2018. 5, 6

[20] Jaihyun Koh, Jangho Lee, and Sungroh Yoon. Single-image deblurring with neural networks: A comparative survey. *Computer Vision and Image Understanding*, 203:103134, 2021. 4

[21] Ashutosh Kulkarni, Prashant W Patil, Subrahmanyam Murala, and Sunil Gupta. Unified multi-weather visibility restoration. *IEEE TMM*, 2022. 3

[22] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *CVPR*, pages 17452–17462, 2022. 3

[23] Heng Li, Xiangping Wu, Qingcai Chen, and Qianjin Xiang. Foreground and text-lines aware document image rectification. In *CVPR*, pages 19574–19583, 2023. 1, 3, 5, 6

[24] Ruoteng Li, Robby T Tan, and Loong-Fah Cheong. All in one bad weather removal using architectural search. In *CVPR*, pages 3175–3185, 2020. 1, 2, 3

[25] 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 *CVPR*, pages 12449–12458, 2023. 2, 6

[26] Yun-Hsuan Lin, Wen-Chin Chen, and Yung-Yu Chuang. Bedsr-net: A deep shadow removal network from a single document image. In *CVPR*, pages 12905–12914, 2020. 2, 3

[27] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. *NeurIPS*, 31, 2018. 3

[28] Wenjie Liu, Bingshu Wang, Jiangbin Zheng, and Wenmin Wang. Shadow removal of text document images using background estimation and adaptive text enhancement. In *ICASSP*, pages 1–5, 2023. 3

[29] Yihao Liu, Xiangyu Chen, Xianzheng Ma, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Unifying image processing as visual prompting question answering. *arXiv preprint arXiv:2310.10513*, 2023. 2, 3

[30] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-IO: A unified model for vision, language, and multi-modal tasks. In *ICLR*, 2022. 1, 3

[31] Weiqi Luo, Jiwu Huang, and Guoping Qiu. JPEG error analysis and its applications to digital image forensics. *IEEE TIFS*, 5(3):480–491, 2010. 8

[32] Jiaqi Ma, Tianheng Cheng, Guoli Wang, Qian Zhang, Xinggang Wang, and Lefei Zhang. Prores: Exploring degradation-aware visual prompt for universal image restoration. *arXiv preprint arXiv:2306.13653*, 2023. 2, 3

[33] Ke Ma, Zhixin Shu, Xue Bai, Jue Wang, and Dimitris Samaras. DocUNet: Document image unwarping via a stacked U-Net. In *CVPR*, pages 4700–4709, 2018. 1, 2, 5, 6

[34] Ke Ma, Sagnik Das, Zhixin Shu, and Dimitris Samaras. Learning from documents in the wild to improve document unwarping. In *ACM SIGGRAPH*, pages 1–9, 2022. 5

[35] Yuhi Matsuo, Naofumi Akimoto, and Yoshimitsu Aoki. Document shadow removal with foreground detection learning from fully synthetic images. In *ICIP*, pages 1656–1660, 2022. 5

[36] Hossein Ziaei Nafchi, Seyed Morteza Ayatollahi, Reza Farrahi Moghaddam, and Mohamed Cheriet. An efficient ground truthing tool for binarization of historical manuscripts. In *ICDAR*, pages 807–811, 2013. 5

[37] Konstantinos Ntirogiannis, Basilis Gatos, and Ioannis Pratikakis. ICFHR 2014 competition on handwritten document image binarization (H-DIBCO 2014). In *ICFHR*, pages 809–813, 2014. 5

[38] Dezhi Peng, Zhenhua Yang, Jiaxin Zhang, Chongyu Liu, Yongxin Shi, Kai Ding, Fengjun Guo, and Lianwen Jin. UPOCR: Towards unified pixel-level ocr interface. *arXiv preprint arXiv:2312.02694*, 2023. 1

[39] Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. PromptIR: Prompting for all-in-one blind image restoration. *arXiv preprint arXiv:2306.13090*, 2023. 3

[40] Ioannis Pratikakis, Basilis Gatos, and Konstantinos Ntirogiannis. H-DIBCO 2010-handwritten document image binarization competition. In *ICFHR*, pages 727–732, 2010. 5

[41] Ioannis Pratikakis, Basilios Gatos, and Konstantinos Ntirogiannis. ICDAR 2011 document image binarization contest (DIBCO 2011). In *ICDAR*, 2011.

[42] Ioannis Pratikakis, Basilis Gatos, and Konstantinos Ntirogiannis. ICFHR 2012 competition on handwritten document image binarization (H-DIBCO 2012). In *ICFHR*, pages 817–822, 2012.

[43] Ioannis Pratikakis, Basilis Gatos, and Konstantinos Ntirogiannis. ICDAR 2013 document image binarization contest (DIBCO 2013). In *ICDAR*, pages 1471–1476, 2013.

[44] Ioannis Pratikakis, Konstantinos Zagoris, George Barlas, and Basilis Gatos. ICFHR 2016 handwritten document image binarization contest (H-DIBCO 2016). In *ICFHR*, pages 619–623, 2016.

[45] Ioannis Pratikakis, Konstantinos Zagoris, George Barlas, and Basilis Gatos. ICDAR2017 competition on document image binarization (DIBCO 2017). In *ICDAR*, pages 1395–1403, 2017. 5

[46] Ioannis Pratikakis, Konstantinos Zagori, Panagiotis Kaddas, and Basilis Gatos. ICFHR 2018 competition on handwritten document image binarization (H-DIBCO 2018). In *ICFHR*, 2018. 5, 6

[47] Ioannis Pratikakis, Konstantinos Zagoris, Xenofon Karagiannis, Lazaros Tsochatzidis, Tanmoy Mondal, and Isabelle Marthot-Santaniello. ICDAR 2019 competition on document image binarization (DIBCO 2019). In *ICDAR*, 2019. 5

[48] Chenfan Qu, Chongyu Liu, Yuliang Liu, Xinhong Chen, Dezhi Peng, Fengjun Guo, and Lianwen Jin. Towards robust tampered text detection in document image: New dataset and new solution. In *CVPR*, pages 5937–5946, 2023. 8

[49] Jaakko Sauvola and Matti Pietikäinen. Adaptive document image binarization. *Pattern recognition*, 33(2):225–236, 2000. 4

[50] Mohamed Ali Souibgui and Yousri Kessentini. DE-GAN: A conditional generative adversarial network for document enhancement. *IEEE TPAMI*, 44(3):1180–1191, 2020. 2, 6, 8

[51] Mohamed Ali Souibgui, Sanket Biswas, Andres Mafla, Ali Furkan Biten, Alicia Fornés, Yousri Kessentini, Josep Lladós, Lluis Gomez, and Dimosthenis Karatzas. TextDIAE: a self-supervised degradation invariant autoencoder for text recognition and document enhancement. In *AAAI*, pages 2330–2338, 2023. 2

[52] Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. Unifying vision, text, and layout for universal document processing. In *CVPR*, pages 19254–19264, 2023. 2, 3

[53] Chris Tensmeyer and Tony Martinez. Document image binarization with fully convolutional neural networks. In *ICDAR*, pages 99–104, 2017. 4

[54] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M Patel. TransWeather: Transformer-based restoration of images degraded by adverse weather conditions. In *CVPR*, pages 2353–2363, 2022. 2, 3

[55] Bingshu Wang and CL Philip Chen. Local water-filling algorithm for shadow detection and removal of document images. *Sensors*, 20(23):6929, 2020. 5, 6

[56] Hongyi Wang, Yang Xue, Jiaxin Zhang, and Lianwen Jin. Scene table structure recognition with segmentation collaboration and alignment. *Pattern Recognition Letters*, 165:146–153, 2023. 2

[57] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, pages 23318–23340, 2022. 1, 2, 3

[58] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *CVPR*, pages 6830–6839, 2023. 1, 2, 3

[59] Yonghui Wang, Wengang Zhou, Zhenbo Lu, and Houqiang Li. UDoc-GAN: Unpaired document illumination correction with background light prior. In *ACM MM*, pages 5074–5082, 2022. 1, 2, 4, 6

[60] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *ACSSC*, pages 1398–1402, 2003. 5

[61] Zongyuan Yang, Baolin Liu, Yongping Xxiong, Lan Yi, Guibin Wu, Xiaojun Tang, Ziqi Liu, Junjie Zhou, and Xing Zhang. DocDiff: Document enhancement via residual diffusion models. In *ACM MM*, pages 2795–2806, 2023. 2, 5, 6, 8

[62] Zongyuan Yang, Baolin Liu, Yongping Xiong, and Guibin Wu. GDB: Gated convolutions-based document binarization. *Pattern Recognition*, 146:109989, 2024. 1, 2, 4, 5, 6

[63] Jaeyoung Yoo, Sang-ho Lee, and Nojun Kwak. Image restoration by estimating frequency distribution of local patches. In *CVPR*, pages 6684–6692, 2018. 8

[64] Shaodi You, Yasuyuki Matsushita, Sudipta Sinha, Yusuke Bou, and Katsushi Ikeuchi. Multiview rectification of folded documents. *IEEE TPAMI*, 40(2):505–511, 2017. 5

[65] 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 *CVPR*, pages 5728–5739, 2022. 5, 6

[66] Francisco Zamora-Martínez, Salvador Espaa Boquera, and María José Castro Bleda. Behaviour-based clustering of neural networks applied to document enhancement. In *IWANN*, 2007. 5

[67] Cheng Zhang, Yu Zhu, Qingsen Yan, Jinqiu Sun, and Yanning Zhang. All-in-one multi-degradation image restoration network via hierarchical degradation representation. In *ACM MM*, pages 2285–2293, 2023. 2, 3

[68] Jiaxin Zhang, Canjie Luo, Lianwen Jin, Fengjun Guo, and Kai Ding. Marior: Margin removal and iterative content rectification for document dewarping in the wild. In *ACM MM*, pages 2805–2815, 2022. 2, 3, 5

[69] Jiaxin Zhang, Bangdong Chen, Hiuyi Cheng, Lianwen Jin, Fengjun Guo, and Kai Ding. DocAligner: Annotating real-world photographic document images by simply taking pictures. *arXiv preprint arXiv:2306.05749*, 2023. 5

[70] Jiaxin Zhang, Lingyu Liang, Kai Ding, Fengjun Guo, and Lianwen Jin. Appearance enhancement for camera-captured document images in the wild. *IEEE Transactions on Artificial Intelligence*, 2023. 1, 2, 4, 5, 6

[71] Ling Zhang, Yinghao He, Qing Zhang, Zheng Liu, Xiaolong Zhang, and Chunxia Xiao. Document image shadow removal guided by color-aware background. In *CVPR*, pages 1818–1827, 2023. 2, 3, 5, 6