ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10889411

Vision Mamba-Based Approach for Incomplete Boundary Document Image Rectification

Weihao Zhang
HiThink RoyalFlush Information Network Co., Ltd.
Hangzhou, China
University of Science and Technology of China
Hefei, China
zhangweihao@myhexin.com

Xin Xia*
HiThink RoyalFlush Information Network Co., Ltd.
Hangzhou, China
xiaxin@myhexin.com

Maopeng Li
HiThink RoyalFlush Information Network Co., Ltd.
Hangzhou, China
limaopeng@myhexin.com

Yunbo Zhao*
University of Science and Technology of China
Hefei, China
ybzhao@ustc.edu.cn

***Abstract*—Capturing document images using handheld mobile devices often results in geometric deformations, which adversely affect the accuracy of Optical Character Recognition (OCR) and document understanding. However, existing transformer-based methods face significant computational costs when processing document images on resource-constrained devices. This study proposes an enhanced Vision Mamba architecture to learn the structural information of document images, thereby rectifying deformed images while reducing computational resource consumption. Additionally, owing to the relative positioning between the document and the imaging device, captured images may exhibit incomplete boundaries. Conventional learning-based methods are primarily designed for images with complete boundaries, which can diminish correction effectiveness. To address this issue, mask consistency loss and preprocessing techniques are introduced to improve the rectification of document images with incomplete boundaries. Experimental results demonstrate the effectiveness and superiority of this method, highlighting its significant value for intelligent document processing.**

***Index Terms*—Document image rectification, vision mamba, image deformation.**

I. Introduction

In our digitally driven society, the transformation of paper documents into electronic formats has become a routine task, particularly through the use of smartphones and portable cameras. This shift offers significant advantages in terms of flexibility and convenience compared to traditional flatbed scanners. However, the process of capturing document image signals with mobile devices is prone to various distortions owing to inherent factors such as the angle of capture, physical deformations of the document, and inconsistent lighting conditions. These distortions can result in issues like perspective skew, uneven illumination, and inclusion of extraneous background elements, which complicate downstream tasks such as OCR, document layout analysis, and content extraction.

Recent advancements in document image rectification have transitioned from traditional 3D reconstruction techniques [1] to deep learning-based methods [2]. Traditional solutions often rely on geometric properties and auxiliary hardware [3] for 3D shape recovery, but these approaches are computationally intensive and impractical for real-world applications. In contrast, deep learning methods, such as those proposed by Ma et al. [4] and Das et al. [5], leverage neural networks to predict pixel displacement fields for the rectification of distorted images. Despite improvements in speed, these methods still face issues of slow processing and high computational costs on resource-constrained mobile devices.

In addition, existing methods [6], [7] primarily focus on improving the rectification performance of document images with complete boundaries, and they struggle to handle cases with incomplete boundaries effectively. In practical applications, it is inevitable to encounter situations where corners and boundaries are missing or occluded. When the distorted image does not contain complete document boundaries, the rectification quality significantly deteriorates.

The contributions of this study are as follows:

1) To address the challenges of high computational complexity and large parameter sizes faced by Transformer models when processing deformed document images on mobile devices, an improved Vision Mamba encoder is used to solve the problems of slow inference speed and high resource consumption in the actual document image rectification process.

2) To improve the rectification performance of document images with boundary occlusion and missing corners in natural scenes, a preprocessing method and a mask consistency loss function are designed to enhance the ability of the model to handle images with incomplete boundaries.

II. Related Work

A. Rectification Based on Deformation Feature Modeling

Document image rectification methods based on deformation feature modeling have been demonstrated to exhibit

*Corresponding author.

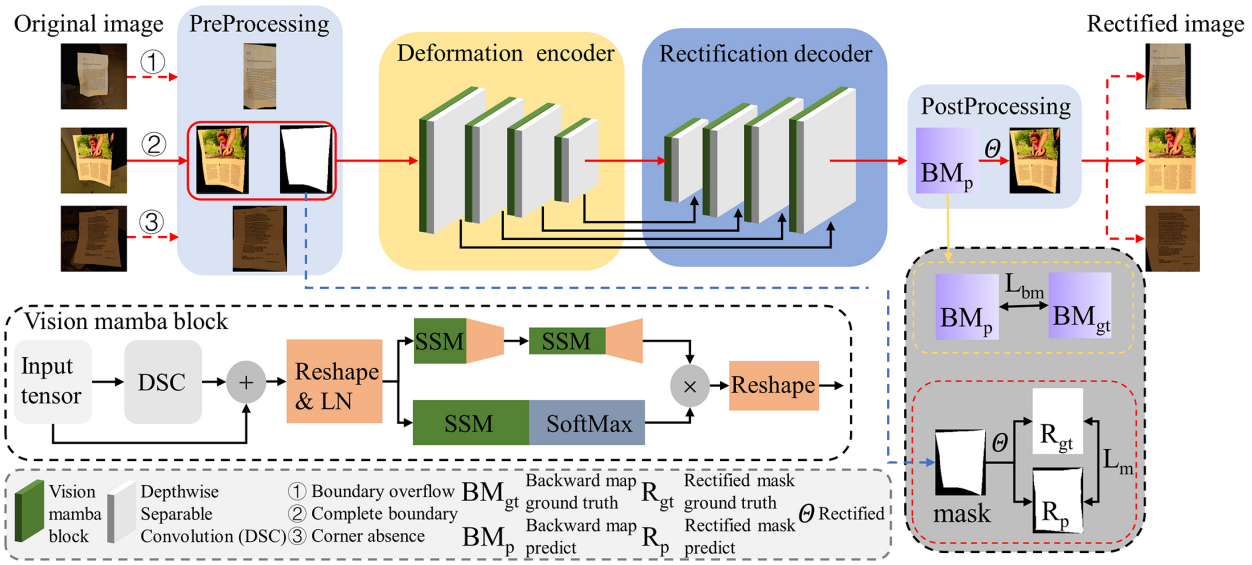

Fig. 1. Overview of the proposed architecture.

strong performance in specific application scenarios. DewarpNet introduced a deep-learning approach leveraging 3D geometry modeling for document image unwarping, leading to substantial OCR performance improvements [5]. A novel approach, UVDoc, employed a pseudo-photorealistic dataset to train a grid-based unwarping network [8]. DocReal utilized an attention-enhanced control point module and background noise removal to improve dewarping accuracy and robustness [2]. DRNet introduced a coarse-to-fine rectification network, leveraging geometric consistency for enhanced results [9].

Despite advancements, most algorithms require complete document boundaries for optimal rectification, and their performance degrades with missing or incomplete boundaries [10]. DocProj [11] addresses this by segmenting images into patches, but stitching them incurs computational overhead and limits applicability to images without backgrounds. MataDoc addressed the challenge of incomplete document boundaries by incorporating margin and text-aware regularizations [12]. However, these Transformer-based models still necessitate high-end GPUs with substantial memory capacity for training and mobile deployment, significantly constraining their widespread applicability.

B. Mamba

State Space Models (SSM) have gained significant attention as potential alternatives to self-attention-based Transformer models [13]. Mamba is a novel SSM architecture designed to capture long-range dependencies while maintaining high performance and reducing computational costs [14]. Unlike traditional Transformer models, Mamba introduces linear state-space equations for modeling dynamic systems. This approach has proven effective across various domains, including natural language processing, computer vision, graph networks, multimodal and multimedia data processing, and time series data [15]–[17].

III. Proposed Framework

The Mamba network, comprising multiple layers of Vision Mamba and feedforward network architecture, captures local features and long-range information by learning the structural information of deformed images, thereby predicting the two-dimensional dense grid coordinates: the backward mapping (BM). Subsequently, in the post-processing stage, BM is utilized for the pixel-wise rectification of the image deformation. The proposed framework is shown in Fig. 1.

A. Preliminaries

1) SSM. The SSM draw inspiration from continuous systems, such as those found in classical control theory [17]. These models transform a one-dimensional input signal $z(t) \in \mathbb{R}$ into an output $w(t) \in \mathbb{R}$ via an intermediate latent variable $q(t) \in \mathbb{R}^N$. The transformation is governed by the parameter matrices $\mathbf{M} \in \mathbb{R}^{N\times N}$, $\mathbf{N} \in \mathbb{R}^{N\times 1}$, and $\mathbf{P} \in \mathbb{R}^{1\times N}$. The state evolution and output generation can be described by the following equations:

$$\begin{aligned} q'(t) &= \mathbf{M}q(t) + \mathbf{N}z(t), \\ w(t) &= \mathbf{P}q(t). \end{aligned} \tag{1}$$

where $q'(t) = \frac{d}{dt}q(t)$. These formulations enable the SSM to capture intricate temporal dependencies by leveraging the latent state to encapsulate both immediate and long-term information from the input sequence.

2) Vision Mamba. Vision Mamba (Vim) overcomes the computational and memory limitations associated with transformer-based understanding of high-resolution images. The foundational Mamba model, originally designed for one-dimensional sequences, has been adapted for visual data processing through a series of transformations [18]. Initially, a 2-D image $\mathbf{I} \in \mathbb{R}^{H\times W\times C}$ is transformed into flattened 2-D patches $\mathbf{P} \in \mathbb{R}^{K\times(Q^2\cdot C)}$, where (H, W) denotes the dimensions of the input image, C represents the number of channels,

and Q specifies the patch size. Subsequently, the patches $\mathbf{P}$ are linearly projected into vectors of dimension D and augmented with positional embeddings $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{(K+1)\times D}$:

$$\mathbf{U}_0 = [\mathbf{c}_{\text{cls}}; \mathbf{c}_q^1\mathbf{W}; \mathbf{c}_q^2\mathbf{W}; \cdots; \mathbf{c}_q^K\mathbf{W}] + \mathbf{E}_{\text{pos}}, \tag{2}$$

where $\mathbf{c}_q^k$ represents the k-th patch of $\mathbf{I}$, and $\mathbf{W} \in \mathbb{R}^{(Q^2\cdot C)\times D}$ is a learnable projection matrix. A class token is employed to represent the entire patch sequence, denoted as $\mathbf{c}_{\text{cls}}$. The token sequence $(\mathbf{U}_{l-1})$ is processed through the l-th layer of the Vim encoder, yielding the output $\mathbf{U}_l$. The output class token $\mathbf{U}_0^1$ undergoes normalization and is subsequently fed into a multi-layer perceptron (MLP) to produce the final prediction $\hat{\mathbf{y}}$:

$$\begin{aligned} \mathbf{U}_l &= \text{Vim}(\mathbf{U}_{l-1}) + \mathbf{U}_{l-1}, \\ \mathbf{g} &= \text{Norm}(\mathbf{U}_0^1), \\ \hat{\mathbf{y}} &= \text{MLP}(\mathbf{g}). \end{aligned} \tag{3}$$

where L signifies the number of layers, and Norm refers to the normalization layer.

B. Preprocessing

The preprocessing stage comprises two operations: document background removal and the random generation of document images with incomplete boundaries. Firstly, owing to the complexity of document image backgrounds, their reconstruction fails to aid in the learning of document structural information by the model [19]. Consequently, the preprocessing stage employs a pre-trained DeepLabv3 [20] semantic segmentation network to obtain document region masks for the elimination of noisy backgrounds. Secondly, to enhance the diversity of document images with incomplete boundaries, training images undergo random scaling of edge and corner cropping, which simulates documents and receipts with missing edges or corners in real-world office scenarios.

C. Network Architecture

As shown in Fig. 1, the deformation encoder and rectification decoder are constructed using a classic U-shaped architecture. The process of the improved Vim block is as follows. It begins with an input tensor $\mathbf{I} \in \mathbb{R}^{H\times W\times C}$. Initially, this tensor undergoes two depthwise separable convolutions followed by a residual connection, ensuring the shape remains $\mathbf{I}_d \in \mathbb{R}^{H\times W\times C}$. Subsequently, the tensor is reshaped and flattened to a form $\mathbf{I}_r \in \mathbb{R}^{L\times C}$, where L is defined as $H \times W$. Layer normalization is then applied. The subsequent flow splits into two parallel SSM branches. In the first branch, the tensor is processed through two SSM layers of different depths to acquire representations of the deformation characteristics of the distorted document. The second branch involves an SSM layer followed by a softmax operation to generate an attention map. Owing to the diversity of deformation characteristics, the SSM layer parameters of the two branches are independent of each other to enhance the model's generalization capability. The outputs of the two branches are integrated via a product operation. Finally, the combined tensor is reshaped back to its original dimensions $\mathbf{I}_m \in \mathbb{R}^{H\times W\times C}$, matching the input tensor shape.

D. Loss Functions

In this study, the training loss function of the document image rectification model comprises two components: first, the L_1 distance between the predicted BM (also known as the deformation field) grid coordinates BM_p and the ground truth BM_{gt}; second, the mask consistency loss. The backward mapping loss is calculated as follows:

$$L_{bm} = \|\text{BM}_{gt} - \text{BM}_p\|_1 \tag{4}$$

For document images with incomplete boundaries, mask consistency loss is introduced to enhance the learning of document image structural information by the model [12]. By utilizing BM to rectify the deformed mask, the L_1 loss is computed to constrain the document boundaries. This constitutes an auxiliary loss function, formulated as follows:

$$L_m = \|\text{R}_{gt} - \text{R}_p\|_1 \tag{5}$$

where R_p and R_{gt} are the binary images obtained by rectifying the deformed masks using BM_p and BM_{gt}, respectively.

The final training loss function is as follows:

$$L_{train} = L_{bm} + \alpha L_m \tag{6}$$

where α represents the weighting coefficient.

IV. Experiments

A. Datasets and Evaluation Metrics

The classic Doc3D dataset was selected to train the document rectification model, consisting of a total of 100K samples [5]. The performance of this model was evaluated using the DocUNet Benchmark dataset, which comprises 130 distorted documents captured in natural scenes [4]. Multi-scale Structural Similarity (MS-SSIM) and Local Distortion (LD) were employed as metrics for assessing the corrected image similarity, facilitating a comparison with previous research efforts [6]. Additionally, computational overhead metrics, such as model parameters and FPS on a single A100 GPU, were incorporated.

B. Experimental Setup

The entire model architecture was implemented using the PyTorch framework. The input image resolution was set to 288×288, utilizing the AdamW optimizer with an initial learning rate of 0.0001. Additionally, a cosine annealing learning rate scheduler with a warmup period of 4 was employed. The model was trained for 65 epochs with a batch size of 48 on an NVIDIA A100 GPU with 80GB of memory. The α in (6) was set to 0.1.

C. Experimental Results

Qualitative Comparison. As shown in Fig. 2, several mainstream open-source models are compared, with validation data encompassing complete boundary documents (Rows 1 and 2), and boundary-lost documents (Rows 3 and 4). Mainstream models, when learning the structural information of document images with incomplete boundaries, are not constrained by the

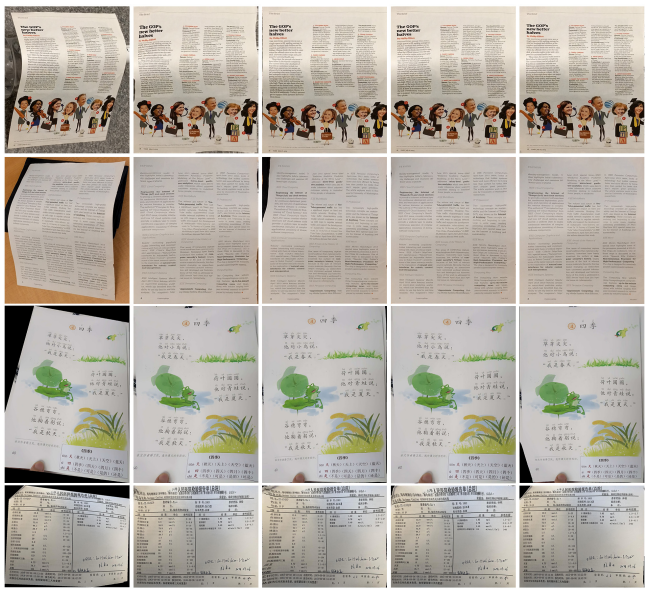

Fig. 2. Qualitative comparison with previous studies. Rows 1 and 2 contain images from the DocUNet benchmark dataset. Rows 3 and 4 show document images with incomplete boundaries.

TABLE I
COMPARISON WITH PREVIOUS STUDIES ON DOCUNET BENCHMARK.

Methods	MS-SSIM↑	LD↓	Params(M)	FPS	Incomplete boundaries
Deformed	0.317	39.58	-	-	-
DocUNet [4] (CVPR2018)	0.410	14.19	69.1	-	×
DewarpNet [5] (ICCV2019)	0.470	8.95	86.9	3.7	×
DocTr [10] (MM2021)	0.510	7.76	26.9	9.5	×
(ICCV2023) [6]	0.497	8.43	63.5	5.1	×
(ICASSP2024) [7]	0.490	7.96	30.2	-	×
No_L_m	0.479	8.16	22.2	13.7	×
Ours	0.498	8.10	22.2	13.7	√

Note: "↑" indicates that higher values are better, while "↓" indicates the opposite. In the FPS column, "-" denotes that the project is not open-source. × indicates the current incapability to rectify documents with incomplete boundaries. All results were obtained on the same A100 GPU.

layout shape, resulting in the forced completion of missing areas during inference. For instance, in the fourth row of Fig. 2, the upper left corner of the rectified image is forcibly filled. This can lead to significant deformation of text and images. The proposed method effectively rectifies typical deformed document images, achieving high accuracy in restoring deformed text and images, resulting in clearer and more recognizable content. Additionally, it demonstrates superior correction for images with incomplete boundaries, enabling it to handle various real-world scenarios, thereby validating the effectiveness of the proposed preprocessing approach.

Quantitative Comparison. Tab. I lists the performance comparison results between the Vim-based model and the Vision Transformer-based models [6], [7], [10] on the DocUNet benchmark dataset. The first row, labeled "Deformed", indicates that the deformed documents were directly evaluated without any rectification methods. As shown in Tab. I, owing to the utilization of the more computationally efficient Vim blocks, our method enhances the processing efficiency of deformed document images, achieving a maximum FPS of 13.7. This enables the Vim-based document rectification model to better serve resource-constrained devices, such as mobile platforms, in restoring deformed text and images. Additionally, the MS-SSIM and LD metrics are 0.498 and 8.10, respectively, reaching mainstream performance levels. This indicates that the improved Vim-based encoder effectively learns long-range contextual information from deformed images while enhancing efficiency. The experimental results demonstrate that the Vim-based model can accelerate inference, and the proposed method exhibits superior computational efficiency compared to most existing methods while maintaining good performance on images with incomplete boundaries.

D. Ablation Study

To verify that mask consistency loss effectively enhances the ability of the model to learn structural information from deformed document images, the following ablation study was conducted. "No_L_m" indicates that mask consistency loss was not utilized during training, with all other settings identical to "Ours." The last two rows of Tab. I present the comparative results of the ablation study. It is evident from the table that the model trained with mask consistency loss improves the correction of document images. This improvement is attributed to the increased focus of the model on the structural information of the layout, particularly enhancing the correction of document images with missing boundaries in real-world scenarios.

V. CONCLUSION

This study explores the optimization and rectification of deformed document image signals captured by mobile devices. Firstly, to address the high computational cost of processing document images on resource-constrained devices, an improved Vim-based encoder is employed to enhance model inference speed. The results indicate that, compared to Transformer-based models, the Vim-based model effectively rectifies deformed images while reducing computational resource consumption. Additionally, the introduction of a mask consistency loss function enhances the performance of the model in handling images with incomplete boundaries, thereby better serving practical application scenarios. In the future, multimodal feature fusion methods will be adopted to optimize the rectification of local distortions in document images, enabling more precise restoration of image detail information.

REFERENCES

[1] G. Meng, Y. Su, Y. Wu, S. Xiang, and C. Pan, "Exploiting vector fields for geometric rectification of distorted document images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 172–187.

[2] F. Yu, Y. Xie, L. Wu, Y. Wen, G. Wang, S. Ren, X. Chen, J. Mao, and W. Li, "Docreal: Robust document dewarping of real-life images via attention-enhanced control point prediction," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 665–674.

[3] G. Meng, Y. Wang, S. Qu, S. Xiang, and C. Pan, "Active flattening of curved document images via two structured beams," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3890–3897.

[4] K. Ma, Z. Shu, X. Bai, J. Wang, and D. Samaras, "Docunet: Document image unwarping via a stacked u-net," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4700–4709.

[5] S. Das, K. Ma, Z. Shu, D. Samaras, and R. Shilkrot, "Dewarpnet: Single-image document unwarping with stacked 3d and 2d regression networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 131–140.

[6] H. Li, X. Wu, Q. Chen, and Q. Xiang, "Foreground and text-lines aware document image rectification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 19 574–19 583.

[7] H. Tang, J. Guo, T. Wang, Y. Yu, and C. Wang, "Efficient joint rectification of photometric and geometric distortions in document images," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 3690–3694.

[8] F. Verhoeven, T. Magne, and O. Sorkine-Hornung, "Uvdoc: Neural grid-based document unwarping," in *SIGGRAPH Asia 2023 Conference Papers*, 2023, pp. 1–11.

[9] S. Liu, H. Feng, and W. Zhou, "Rethinking supervision in document unwarping: a self-consistent flow-free approach," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[10] H. Feng, Y. Wang, W. Zhou, J. Deng, and H. Li, "Doctr: Document image transformer for geometric unwarping and illumination correction," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, p. 273–281.

[11] X. Li, B. Zhang, J. Liao, and P. V. Sander, "Document rectification and illumination correction using a patch-based cnn," *ACM Transactions on Graphics (TOG)*, pp. 1–11, 2019.

[12] B. Dai, Q. Xie, Y. Li, X. Qin, C. Zhang, K. Yao, J. Han *et al.*, "Matadoc: Margin and text aware document dewarping for arbitrary boundary," *arXiv preprint arXiv:2307.12571*, 2023.

[13] A. Gu, I. Johnson, K. Goel, K. Saab, T. Dao, A. Rudra, and C. Ré, "Combining recurrent, convolutional, and continuous-time models with linear state space layers," *Advances in Neural Information Processing Systems*, pp. 572–585, 2021.

[14] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[15] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, "Vmamba: Visual state space model," *arXiv preprint arXiv:2401.10166*, 2024.

[16] X. Wang, S. Wang, Y. Ding, Y. Li, W. Wu, Y. Rong, W. Kong, J. Huang, S. Li, H. Yang *et al.*, "State space model for new-generation network alternative to transformers: A survey," *arXiv preprint arXiv:2404.09516*, 2024.

[17] R. Xu, S. Yang, Y. Wang, B. Du, and H. Chen, "A survey on vision mamba: Models, applications and challenges," *arXiv preprint arXiv:2404.18861*, 2024.

[18] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *arXiv preprint arXiv:2401.09417*, 2024.

[19] S. Liu, H. Feng, W. Zhou, H. Li, C. Liu, and F. Wu, "Docmae: Document image rectification via self-supervised representation learning," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2023, pp. 1613–1618.

[20] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.