DocAttentionRect: Attention-Guided Document Image Rectification

Pooja Kumari, Sukhendu Das
IIT Madras, India.
jaiswal12pooja@gmail.com, sdas@cse.iitm.ac.in

BMVC 2025

Introduction:

- Design of an efficient system for document image rectification from a single input image. Addresses challenges of image distortions, noise, complex layouts, text segmentation while optimizing OCR performance.
- Existing SoA algorithms fail to achieve satisfactory rectification in presence of incomplete boundaries or with regions containing partial text.
- This paper introduces a novel approach to document image rectification that specifically addresses the issue of handling images with partial or no document boundaries.
- DocAttentionRect demonstrates versatility by handling all types of documents, regardless of their boundary visibility. Extensive experiments validate the effectiveness and state-of-the-art performance.

Proposed Architecture:

- The proposed work operates as a 2-Stage Document Rectification Pipeline, beginning with a Lightweight Segmentation module that performs robust background removal and document isolation.
- The first rectification stage, uses a Control Point based Coarse Rectification (CCR) module for initial geometry correction. This coarsely corrected image is then refined by the Attention-based Rectification Network (ARN), responsible for predicting a precise, pixel-wise Flow Map—a 2-D vector field defining pixel-wise paths for perfect unwrapping.
- Finally, the rectification process concludes with Pixel Evaluation, Sampling, and Unwrapping using this flow map to generate the clean, rectified document image.

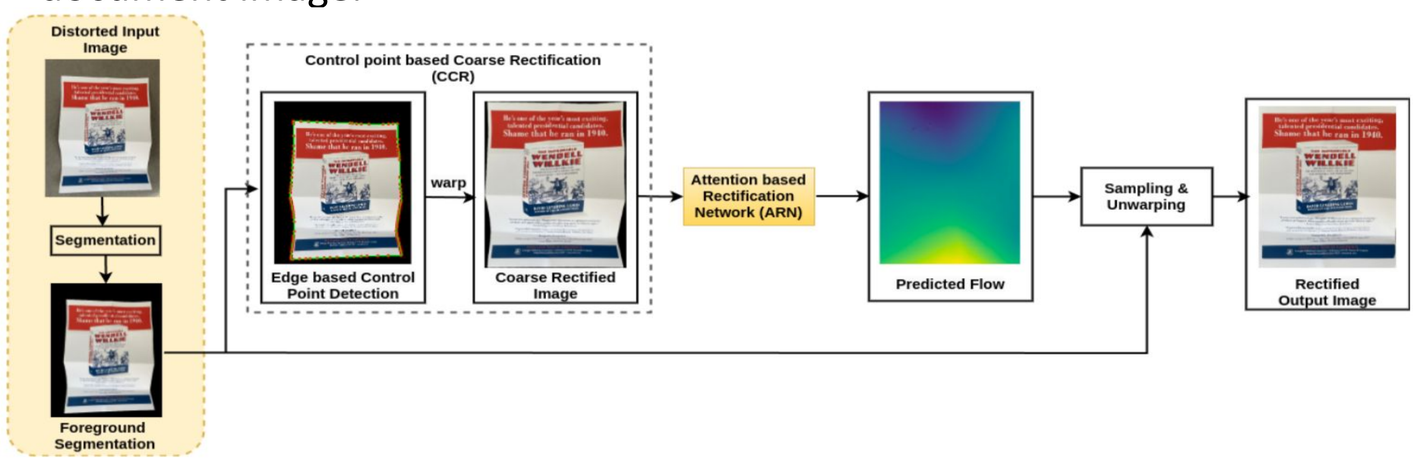

Figure 1: Brief architecture of our proposed model for document image rectification.

Figure 2: Overall steps of Control point based Coarse Rectification (CCR) module. All stages are unsupervised, ending with a perspective correction using point correspondences.

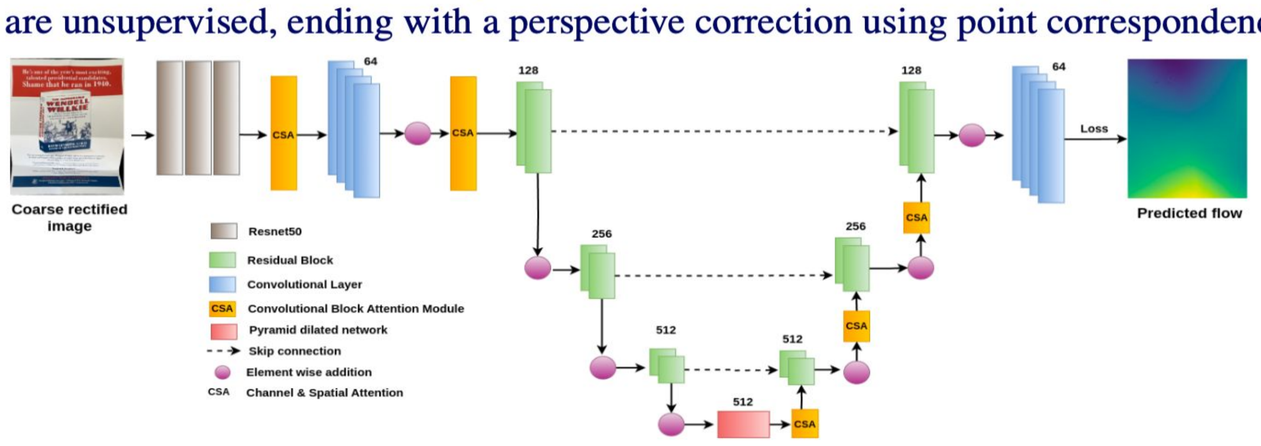

Figure 3: Overview of our proposed Attention-Based Rectification Network (ARN).

Loss Functions:

- **Cross entropy loss**

$$\mathcal{L}_{\text{seg}} = -\sum_{i=1}^{N_P} [y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)]$$

- **Charbonnier loss**

$$\mathcal{L}_{\text{C}} = \sqrt{(p_i - \hat{p}_i)^2 + \epsilon^2}$$

- **Element-wise L1 loss**

$$\mathcal{L}_1 = \lambda_1 . \frac{1}{N_f} \sum_{i}^{N_f} \left\| d_i - \hat{d}_i \right\|_1$$

- **Endpoint Error (EPE) loss**

$$\mathcal{L}_{\text{EPE}} = \lambda_2 . ||v_{est} - v_{gt}||$$

Qualitative Results:

Figure 4: Qualitative comparison of our DocAttentionRect with existing SOTA.

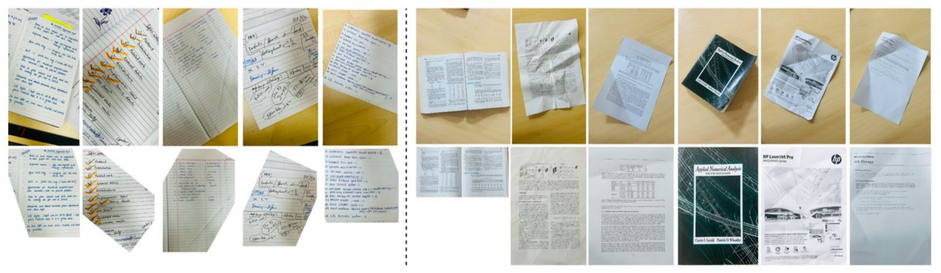

Figure 5: Qualitative results of DocAttentionRect on (left) handwritten documents with missing boundaries and (right) mobile-captured images with wide backgrounds.

Quantitative Results:

Table 1: Quantitative comparison of our work with existing SOTA for the DocUNet dataset.

Methods	Venue	MS-SSIM ↑	LD ↓	ED ↓	CER ↓
Distorted	-	0.25	20.51	1552.22	0.5089
DocUNet	CVPR'18	0.41	14.19	1259.83	0.3966
DocProj	TOG'19	0.29	18.01	1165.93	0.3818
FCN-based	DAS'20	0.45	7.84	1031.40	0.3156
DewarpNet	ICCV'19	0.47	8.39	525.45	0.2102
DocTr [3]	MM'21	0.51	7.76	464.83	0.1746
PWUNet	ICCV'21	0.49	8.64	743.32	0.2623
DDCP	ICDAR'21	0.47	8.99	745.35	0.2626
Marior	MM'22	0.48	7.44	593.80	0.2136
RDGR	CVPR'22	0.50	8.51	420.25	0.1559
DocGeoNet [2]	ECCV'22	0.50	7.71	379.00	0.1509
DocTr++	IEEE TMM'23	0.51	7.52	447.47	0.1695
CGU-Net	Arxiv'23	0.51	7.18	689.54	0.1840
MataDoc	Arxiv'23	0.50	7.42	392.48	0.1505
Foreground and Text	ICCV'23	0.49	8.43	697.52	0.1705
LA-DocFlatten	TOG'23	**0.52**	6.72	391.90	0.1530
DRNet	IEEE TCSVT'23	0.51	7.42	644.48	0.1649
UVDoc	SIGGRAPH Asia'23	0.54	6.83	707.00	0.1720
PolarDoc	Arxiv'23	0.49	7.24	868.88	0.2460
DocTLNet	IJDAR'24	0.51	**6.70**	377.12	0.1504
DocReal [1]	WACV'24	0.50	7.03	359.90	0.1441
Ours (DocAttentionRect)	-	**0.52**	**6.70**	**356.21**	**0.1432**

Table 2: Quantitative comparison with other existing methods for the DIR300 test set.

Methods	Venue	MS-SSIM ↑	LD ↓	ED ↓	CER ↓
Distorted	-	0.3169	39.58	1500.56	0.5234
DocProj	TOG'19	0.3246	30.63	958.89	0.3540
DewarpNet	ICCV'19	0.4921	13.94	1059.57	0.3557
DocTr [3]	MM'21	0.6160	7.21	699.63	0.2237
DDCP	ICDAR'21	0.5524	10.95	2084.97	0.5410
DocGeoNet [2]	ECCV'22	0.6380	6.40	664.96	0.2189
Foreground ...	ICCV'23	0.6070	7.68	652.80	0.2115
MataDoc	Arxiv'23	0.6389	5.75	482.90	0.1785
DocTLNet	IJDAR'24	**0.6582**	5.75	482.57	0.1767
PolarDoc	Arxiv'23	0.6050	7.17	549.64	0.1680
Ours (DocAttentionRect)	-	0.6401	**5.73**	**482.51**	**0.1700**

Conclusion:

- Our study addresses the challenge of document image rectification by introducing a novel approach that specifically targets images with partial or absent document boundaries.
- By leveraging attention mechanisms and parallel convolution layers, DocAttentionRect accurately captures long-range dependencies and crucial textual and structural components throughout the rectification process.
- Extensive experimentation on 3 benchmark datasets demonstrates the effectiveness and state-of-the-art performance.

References:

1. Fangchen Yu, Yina Xie, Lei Wu, Yafei Wen, Guozhi Wang, Shuai Ren, Xiaoxin Chen, Jianfeng Mao, and Wenye 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, pages 665–674, 2024.
2. Hao Feng, Wengang Zhou, Jiajun Deng, Yuechen Wang, and Houqiang Li. Geometric representation learning for document image rectification. In European Conference on Computer Vision (ECCV), pages 475–492, 2022.
3. Feng, Hao, Yuechen Wang, Wengang Zhou, Jiajun Deng, and Houqiang Li. "Doctr: Document image transformer for geometric unwarping and illumination correction." arXiv preprint arXiv:2110.12942, 2021.