# DocAttentionRect: Attention-Guided Document Image Rectification

Pooja Kumari
jaiswal12pooja@gmail.com
Sukhendu Das
sdas@iitm.ac.in
Indian Institute of Technology Madras, Chennai, India.

**Abstract**

In recent years, document image rectification has seen substantial advancements. Nonetheless, current leading algorithms are primarily effective for images with clearly defined document boundaries and some degree of distortion. These algorithms often struggle when presented with images containing text in only a specific area or with incomplete boundaries, leading to subpar rectification results. This limitation is particularly problematic in situations where only sections of a document need processing. Although there are methods that attempt to address these issues, they frequently encounter difficulties when dealing with a combination of intricate distortions and diverse document layouts. To address this gap, our paper introduces a novel approach for document image rectification that specifically targets images with partial or missing document boundaries. Recently, attention-based neural networks have proven highly effective in enhancing the accuracy and efficiency of document rectification. By utilizing attention mechanisms, these networks can focus on relevant parts of an image, thereby improving the rectification outcomes. Our paper presents 'DocAttentionRect', an innovative attention-based rectification network that incorporates attention modules alongside parallel convolution layers to address complex document image rectification challenges. Our proposed architecture captures extensive dependencies and key textual and structural features throughout the rectification process. DocAttentionRect is capable of handling all document types, regardless of the visibility of their boundaries. Extensive experiments conducted on the DocUNet, DIR300, and UVDoc datasets demonstrate the superior performance and effectiveness of our proposed architecture.

## 1 Introduction

With the growing use of smartphones for document digitization, issues like camera angle and paper deformation often introduce distortions in captured images. These distortions hinder readability and affect downstream tasks such as OCR, retrieval, and editing. To address this challenge, researchers have begun exploring unconstrained document image rectification as a promising research area. While traditional methods based on 3D reconstruction attempt to correct such distortions, they usually require multi-view images or specialized hardware, limiting their practicality. Recently, deep learning-based solutions have surfaced as potential alternatives to conventional approaches. These solutions involve resampling the pixels and

© 2025. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

employing trained networks to predict the warping flow explicitly, enabling the repair of warped document images. Additionally, attention-based neural networks have emerged as a powerful tool for addressing a variety of computer vision tasks. Attention-based models show strong potential by focusing on important image regions, leading to more accurate and efficient rectification.

In this paper, we introduce DocAttentionRect, a novel attention-driven framework for document image rectification. Unlike conventional methods that depend on handcrafted features or rigid geometric models, our approach leverages attention mechanisms to learn rectification directly from data, enabling it to adapt effectively to various document layouts and distortion types. The integration of attention is central to our architecture; it empowers the model to focus computational effort on the most relevant regions of the image, enhancing correction in areas that need it most. This dynamic attention also enables the capture of long-range dependencies and structural context, resulting in more accurate and consistent rectifications. Additionally, we incorporate a Control Point-based Coarse Rectification (CCR) module that utilizes document edge cues to generate an initial global alignment. Furthermore, we embed parallel convolutional operations within the network, which not only improve processing speed but also ensure better global feature extraction across the document. To sum up, we have made four-fold contributions, which are as follows:

- We propose DocAttentionRect, a new attention-based framework for document rectification that integrates attention with parallel convolutions.
- The architecture effectively captures global and local features, enhancing structural understanding across various distortion types.
- Our proposed method is able to handle all types of documents, including those with clearly defined boundaries, partially visible boundaries, and no discernible boundaries.
- Extensive experiments on three benchmark datasets demonstrate the state-of-the-art performance of our method.

## 2 Related Works

The document image rectification task has garnered significant attention from both the computer vision and document image processing fields. Recent studies in this domain can be broadly classified into two main categories:

- **Document image rectification using 3D shape reconstruction:** Many earlier techniques rely on additional hardware or prior knowledge of input data to create a three-dimensional representation of the document. For instance, Brown and Seales [3] utilized an illumination projector system to capture the document's three-dimensional shape, followed by flattening using a 'mass-spring particle' system. A laser range scanner was employed by Zhang et al. [24] to construct the document's 3D layout, utilizing physical modeling techniques for restoration. Meng et al. [26] employed structural beams to restore spatial curvatures. Some approaches involve generating the document's 3D shape from multiple image perspectives, as seen in works by Tsoi et al. [33]. Additionally, there are studies focused on reconstructing 3D shapes from a single image for rectification, such as those by Courteile et al. [4] utilizing shape-from-shading techniques.

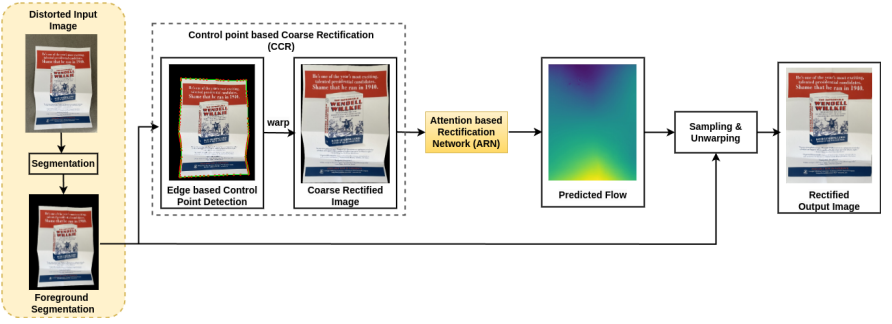


Figure 1: Brief architecture of our proposed model for document image rectification.

- **Document image rectification using deep learning techniques:** The second category of approaches involves data-driven strategies, where deep neural networks are increasingly utilized for document image rectification. For example, DocUNet [24] employs a learning-based approach to estimate pixel-wise displacement fields for correcting document image distortions. Li et al. [20] proposed a method that involves creating patches from the image, rectifying them, and then stitching them together for overall rectification. Das et al. [6] propose DewarpNet, which learns 3D shape of documents using a UNet architecture [31]. Additionally, Feng et al. [9] introduced a transformer-based method for document rectification. DocTr++ [11] employs a hierarchical encoder-decoder architecture to extract multi-scale representations and establish pixel-wise mapping relationships between distorted document images and their undistorted counterparts. Another work, DocReal [42] utilizes Enet to eliminate background noise and integrates an attention-enhanced control point module to enhance the capture of local deformations.

# 3 Methodology of Our Proposed Work

Document images often exhibit irregular structures due to distortions that warp the originally planar text layout. Our approach uses an attention mechanism combined with parallel convolution to effectively capture both local and global distortions, focusing on key regions to improve rectification. Figure 1 illustrates our framework, starting with a distorted input image $I_d$. We first apply a lightweight segmentation module [29] to remove the background, followed by an unsupervised Control Point-based Coarse Rectification (CCR) to detect boundary control points and produce a rough rectified image $I_r$. This image is then refined by the Attention-based Rectification Network (ARN).

Our network comprises two main components: Control point based Coarse Rectification (CCR) and Attention-based Rectification Network (ARN). In a supervised learning environment, our proposed network, DocAttentionRect, learns the transformation from the image domain $I_d$ to the flow domain $F$. The flow, represented as a two-dimensional vector field, delineates the path that pixels should follow from the input image to create the corrected image. To achieve the desired outcome, a resampling technique is necessary, as the flow essentially maps from the distorted image to the corrected/rectified image in the forward direction. This forward mapping also indicates the locations of all known-color pixels in the

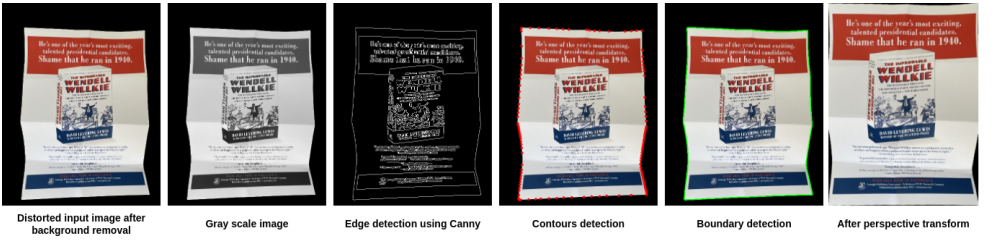


Figure 2: Overall steps of Control point based Coarse Rectification (CCR) module. All stages are unsupervised, ending with a perspective correction using point correspondences.

warped image. To reconstruct the final undistorted image based on the distortion flow, we employ resampling of the given image and utilize a pixel evaluation algorithm [20] to find the mapping. The bidirectional iterative search algorithm [41] serves as inspiration for this method. The objective of the conventional backward mapping algorithm [41] is to search for a point $q$ in the input image that maps to $r$ in the output image. This method effectively reverses the mapping using a continuous search until point $q$ converges, given that we only have the forward distortion flow, as represented by the below equations:

$$q^{(0)} = r \tag{1}$$

$$q^{(i+1)} = r \quad f \quad q^{(i)} \tag{2}$$

where $f(q)$ denotes the predicted forward flow from input pixel $q$ to the rectified output.

## 3.1 Pre-processing

The distorted input image $I_d \; 2 \; \mathbb{R}^{H \quad W \quad 3}$ contains background textures that must be removed before rectification. Almost all existing approaches incorporate some form of preprocessing, either as a standalone module or integrated within their primary architecture. Our approach includes a preprocessing step using a lightweight segmentation network [29] to generate a confidence map. This map is binarized with an empirically set threshold to extract the document foreground. The background is removed by element-wise multiplying this mask with the original image. This preprocessing module is trained separately using binary cross-entropy loss [8], defined as:

$$L_{\text{seg}} = \mathop{\text{å}}_{i=1}^{N_q} [y_i \log(\hat{q}_i) + (1 \quad y_i) \log(1 \quad \hat{q}_i)] \tag{3}$$

Here, $y_i$ and $\hat{q}_i$ represent the ground truth and predicted confidence maps, respectively, and $N_q = H \quad W$ is the total number of pixels.

## 3.2 Control point based Coarse Rectification (CCR)

Document images often suffer from global misalignment due to camera angles and paper deformation. To address this, we use edge cues in the image. First, Canny edge detection [30] extracts edges from the distorted input. We then identify contours—curves joining points with similar intensity, and assume, based on common document traits, that the largest

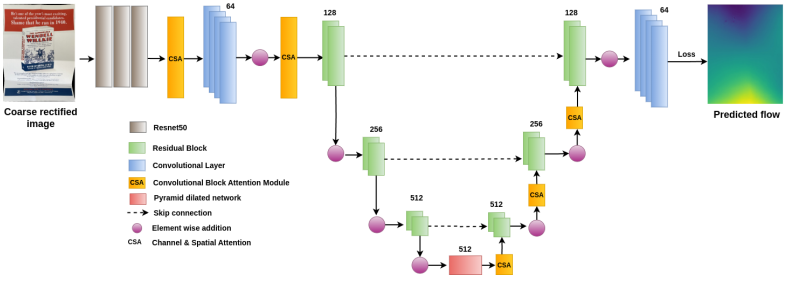


Figure 3: Overview of our proposed Attention-Based Rectification Network (ARN).

contour corresponds to the document boundary. Using this boundary, we apply a perspective transform [12] to generate a coarsely rectified image with corrected global structure, as shown in Figure 2. Additional details are provided in the supplementary.

## 3.3 Attention based Rectification Network (ARN)

The proposed ARN module, as seen in Figure 3, follows an encoder-decoder design tailored to address both local and global distortions. It refines the coarse output from the CCR using a series of ResNet50 blocks, followed by a Convolutional Block Attention Module (CBAM) [33] and subsequent convolutional layers. The attention mechanism directs the network's focus towards the most crucial aspects of the input data, amplifying the importance of certain parts while reducing the significance of others. By integrating spatial and channel attention mechanisms, the module facilitates the adaptive modulation of feature map weights. This attention mechanism enhances the ability of CNNs to capture complex visual patterns and dependencies, resulting in more resilient feature representations. Following each convolutional layer, the model incorporates group normalization and ReLU activation, maintaining an auto-encoder structure. The encoder utilizes four parallel convolutional layers with the kernel sizes of 7x7 and strides of 2 to extract local features efficiently.

Parallel convolution allows the simultaneous extraction of the features from different regions of the input image, facilitating to capture both local and global information. This enhances the robustness of feature representations, thus improving the overall effectiveness of the model. Furthermore, the concurrent processing of data by parallel convolutional layers enables the network to handle the larger datasets without a significant increase in computational complexity. Additionally, parallel convolutions introduce multiple layers of non-linearity through distinct activation functions applied to each convolutional path, potentially capturing more complex data patterns and relationships. Element-wise addition is employed to combine the outputs of individual parallel convolutions for subsequent fusion, followed by residual blocks and the Pyramid Dilated Network (PDN) [14] in our architecture. The inclusion of dilated convolutions with varying dilation rates in the PDN enables the model to analyze features across a range of receptive fields. This capability enhances the PDN's capacity to comprehend intricate spatial configurations within the document image by capturing information at different scales. Furthermore, the hierarchical pyramid architecture of the PDN supports the extraction of features spanning from fine-grained details to broader semantic concepts. Subsequently, we utilize residual blocks, followed by the CBAM, and finally we transpose convolutional layers to restore the spatial features, effectively recovering the original resolution of the image.

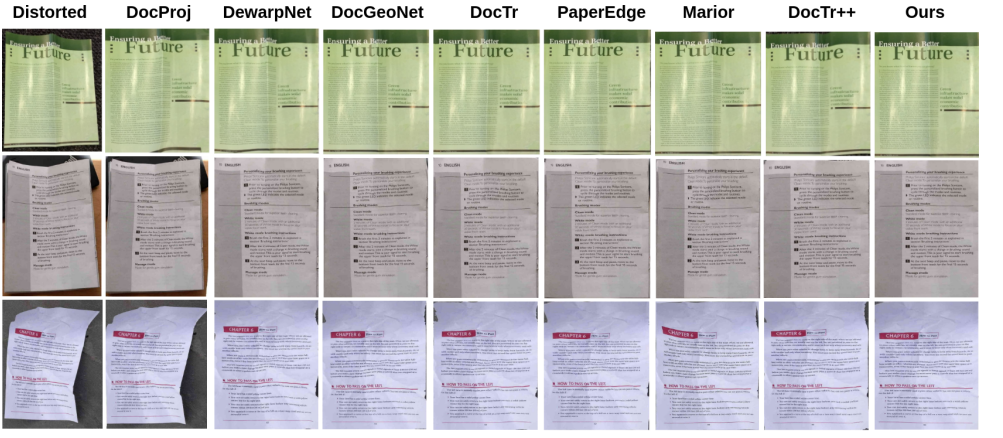


Figure 4: Qualitative comparison of our DocAttentionRect with existing SOTA.

## 3.4 Loss Functions

The proposed model is trained in a supervised manner. Complete model is optimized by reducing the loss function $L$ which is a combination of three loss components.

$$L = L_1 + L_C + L_{EPE} \tag{4}$$

Firstly, we use element-wise L1 loss [28], which measures the difference in foreground pixel displacement between the actual and the predicted flow. Element-wise L1 loss is instrumental in guiding the model to minimize discrepancies at the pixel level, thereby facilitating accurate rectification that preserves the content and structure of the original document.

$$L_1 = l_1 : \frac{1}{N} \sum_{i}^{N} \quad d_i \quad \hat{d}_i \quad _1 \tag{5}$$

where, $N$ is the number of pixels. $d_i$ depicts the ground truth value for the $i^{th}$ data point and $\hat{d}_i$ is the predicted value for the $i^{th}$ data point. Next component is the Charbonnier loss [21] that creates a seamless shift between the L1 and L2 loss functions, thereby effectively managing the balance between the speed of convergence and robustness.

$$L_C = \sqrt{(p_i \quad \hat{p}_i)^2 + e^2} \tag{6}$$

where, $p_i$ represents the predicted pixel value, $\hat{p}_i$ is the corresponding ground truth pixel value, and $e$ is hyper parameter that is added to stabilize the computation. We also used Endpoint Error (EPE) loss [2] which plays a vital role in document image rectification by evaluating the precision of predicted displacement fields against the actual data. It calculates the Euclidean distance between predicted and ground truth displacement vectors for individual pixels, offering a comprehensive assessment of rectification performance.

$$L_{EPE} = l_2 : \frac{1}{N} \sum_{i=1}^{N} \| v_{est_i} \quad v_{gt_i} \|_2 \tag{7}$$

where, $v_{est_i}$ is the predicted flow vector for pixel $i$, $v_{gt_i}$ is the corresponding ground truth flow vector for pixel $i$, $N$ is the total number of pixels, and $\| \cdot \|_2$ denotes the Euclidean norm.

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

| Methods | Venue | MS-SSIM " | LD # | ED # | CER # |
|---|---|---|---|---|---|
| Distorted | - | 0.25 | 20.51 | 1552.22 | 0.5089 |
| DocUNet [24] | CVPR'18 | 0.41 | 14.19 | 1259.83 | 0.3966 |
| DocProj [20] | TOG'19 | 0.29 | 18.01 | 1165.93 | 0.3818 |
| FCN-based [39] | DAS'20 | 0.45 | 7.84 | 1031.40 | 0.3156 |
| DewarpNet [6] | ICCV'19 | 0.47 | 8.39 | 525.45 | 0.2102 |
| DocTr [9] | MM'21 | 0.51 | 7.76 | 464.83 | 0.1746 |
| PWUNet [7] | ICCV'21 | 0.49 | 8.64 | 743.32 | 0.2623 |
| DDCP [40] | ICDAR'21 | 0.47 | 8.99 | 745.35 | 0.2626 |
| Marior [43] | MM'22 | 0.48 | 7.44 | 593.80 | 0.2136 |
| RDGR [15] | CVPR'22 | 0.50 | 8.51 | 420.25 | 0.1559 |
| DocGeoNet [10] | ECCV'22 | 0.50 | 7.71 | 379.00 | 0.1509 |
| DocTr++ [11] | IEEE TMM'23 | 0.51 | 7.52 | 447.47 | 0.1695 |
| CGU-Net [34] | Arxiv'23 | 0.51 | 7.18 | 689.54 | 0.1840 |
| MataDoc [5] | Arxiv'23 | 0.50 | 7.42 | 392.48 | 0.1505 |
| Foreground and Text [18] | ICCV'23 | 0.49 | 8.43 | 697.52 | 0.1705 |
| LA-DocFlatten [19] | TOG'23 | 0.52 | 6.72 | 391.90 | 0.1530 |
| DRNet [23] | IEEE TCSVT'23 | 0.51 | 7.42 | 644.48 | 0.1649 |
| UVDoc [35] | SIGGRAPH Asia'23 | 0.54 | 6.83 | 707.00 | 0.1720 |
| PolarDoc [45] | Arxiv'23 | 0.49 | 7.24 | 868.88 | 0.2460 |
| DocTLNet [16] | IJDAR'24 | 0.51 | 6.70 | 377.12 | 0.1504 |
| DocReal [42] | WACV'24 | 0.50 | 7.03 | 359.90 | 0.1441 |
| Ours (DocAttentionRect) | - | 0.52 | 6.70 | 356.21 | 0.1432 |

# 4 Dataset Details

The most prominent datasets for document image recti cation are Doc3D [6] and DocUN [24]. Doc3D offers 100,000 synthetically warped images with precise 3D annotations, se ing as a comprehensive training dataset. DocUNet, in contrast, is used for evaluation includes 130 images from 65 real documents. We also utilize DIR300 [10], a set of 3 real-world document images captured via mobile devices, offering greater variability a complexity. Additionally, the UVDoc dataset [35] provides diverse pseudo-photorealist images—spanning printed text, handwriting, forms, and diagrams—captured under vary real-world conditions, enhancing evaluation robustness.

# 5 Experiments

## 5.1 Metrics

We utilized several evaluation metrics to compare our method with existing SOTA, inclu ing Structural Similarity (SSIM) [37], Edit Distance (ED) [17], Local Distortion (LD) [22], Multi-Scale Structural Similarity (MS-SSIM) [36], Aligned distortion (AD) [25], and Char- acter Error Rate (CER) [13], to assess and validate the effectiveness of our proposed proach. The details of these metrics can be found in the supplementary document.

## 5.2 Implementation Details

DocAttentionRect was implemented using PyTorch framework. We utilized a batch size 30 during the training time, employing the Adam optimizer with an initial learning rate o 1 10 $^{4}$. The learning rate was reduced by a factor of 0.1 after 20 epochs. Training w