Foreground and Text-lines Aware Document Image Rectification

Heng Li[1] Xiangping Wu[1, 3*] Qingcai Chen[1, 2*] Qianjin Xiang[1]

[1]Harbin Institute of Technology Shenzhen, China, [2]PengCheng Laboratory, Shenzhen China
[3]The Hong Kong Polytechnic University, Hung Hom Kowloon, Hong Kong

hengli.lh@outlook.com, wxpleduole@gmail.com, qingcai.chen@hit.edu.cn, randy19991009@outlook.com

Abstract

This paper aims at the distorted document image rectification problem, the objective to eliminate the geometric distortion in the document images and realize document intelligence. Improving the readability of distorted documents is crucial to effectively extract information from deformed images. According to our observations, the foreground and text-line of the original warped image can represent the deformation tendency. However, previous distorted image rectification methods pay little attention to the readability of the warped paper. In this paper, we focus on the foreground and text-line regions of distorted paper and proposes a global and local fusion method to improve the rectification effect of distorted images and enhance the readability of document images. We introduce cross attention to capture the features of the foreground and text-lines in the warped document and effectively fuse them. The proposed method is evaluated quantitatively and qualitatively on the public DocUNet benchmark and DIR300 Dataset, which achieve state-of-the-art performances. Experimental analysis shows the proposed method can well perform overall geometric rectification of distorted images and effectively improve document readability (using the metrics of Character Error Rate and Edit Distance). The code is available at `https://github.com/xiaomore/Document-Image-Dewarping`.

1. Introduction

In our daily life, there are various documents, such as papers, posters, receipts and so on. These documents contain a wealth of information. At present, document intelligence can automatically extract the information in the document, and then obtain structured data (text content, key field extraction, table structure, document layout analysis), which greatly facilitates people's lives. However, with the popularization of mobile electronic devices, it is more and more convenient for people to take pictures with smart phones or portable cameras to capture electronic documents. When taking pictures, due to the different positions, illumination conditions of these devices, the angle of paper placement, deformation and other problems, the images taken are distorted or deformed to varying degrees, and it is difficult to extract useful information from these distorted pictures, as shown in Figure 1. Even wrong information is extracted, which brings great challenges to obtaining key information of documents and realizing document intelligence.

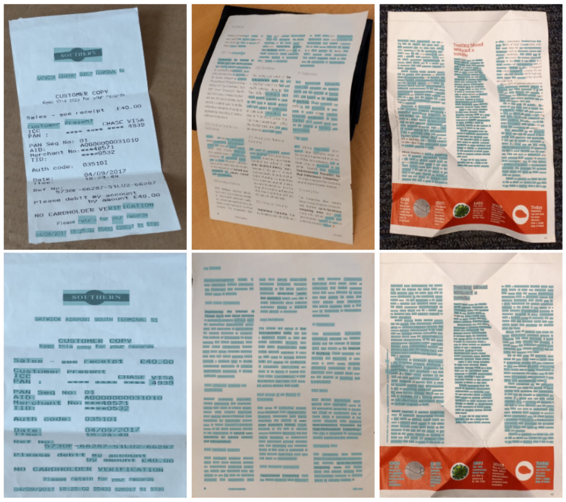

Figure 1. Compared with the text results using Tesseract [32] as OCR engine. Row 1: distorted images. Row 2: rectified by our method. We use color to highlight the detected text boxes.

As early as many years ago, the problem of document image distortion rectification has been paid attention to. In the traditional rectification method before deep learning was widely used, several previous methods [3, 36, 17] according to the deformation characteristics of the image, relying on multi-view, get the best boundary through the regression method, and use the boundary segmentation method to dewarped the segmentation graphics. Tian and Narasimhan

[*] Corresponding authors.

[35] reconstructs warped document images from 2D to 3D by detecting text-lines in the image. In the time since, some works [16, 15] have also captured prior knowledge in deformed images by detecting text lines in order to bring gains to image rectification. However, these methods of restoring the 3D shape of images through auxiliary hardware or multi-view methods will bring time consuming and expensive problems.

With the arrival of the era of deep learning, Convolutional Neural Network (CNN) and Transformer have been introduced to solve the geometric rectification of distorted document images. Most current methods dewarping the input image by learning the mapping relationship between the input distorted image and the dense 2D coordinate map area [26, 8, 1, 11, 14, 10]. The method of [26, 8] used the UNet structure networks to regress the dense 2D coordinates mapping according to the input distorted image. However, it is difficult for CNN to capture the long-distance dependence of distorted paper. For methods [11, 10] that focus on image foreground, they proposed to remove the document background in the preprocessing stage before geometric rectification the network. And then introduces the multi-layer transformer of Encoder-Decoder structure after the convolution network to dewarp the distorted paper. But these methods based on setting the binarization threshold to remove the background will lose the information of the original image. There are also some works such as [14, 10] proposed to use document boundaries, 3D world coordinates and text-lines to construct deformation fields, which reduces the error rate of text recognition. These two methods have no obvious interaction between the text-line and the original distorted image, which makes the text-line area unable to be effectively focused.

In geometrically deformed images, the distortion degree of image foreground and text-lines varies with different scenes. It is necessary to establish an explicit invariant feature in the distorted image foreground to represent the core document and text-line region features robustly, to suppress the interference of background and deformation. Since the foreground information of the document image can clearly represent the deformation amplitude, we use the foreground features representation of the input image. The text-line regions can represent the distorted state of the local position in dewarped images. In order to enhance the reading quality of the image, we considering it is necessary to pay more attention to the area of the text-lines. Therefore, we propose to represent local information using text-line information, extract the characteristics of the corresponding position in the image. Besides, based on our observations, when using the deformation field for distortion rectification, for the dewarped image, the areas in the same horizontal direction should have the same vertical coordinates in the deformation field. For example, for text-lines, the text area of the same line should have the same ordinate. This is a very critical starting point for our design model.

From the perspective of readability, we propose a novel distortion image rectification model based on cross attention mechanism, which fuses the features of the image foreground and text line. The features of the foreground and text-line regions respectively focus on the features of the corresponding positions in the original warped image. The foreground and text-line features are each represent the global and local information of the distorted image, which have complementary effects. In addition, in order to enhance the interaction between global and local features, we share the two cross attention branches of foreground and text-lines, which also reduces the amount of model parameters to a certain extent.

The main contributions of our work are summarized as follows:

- We propose to use foreground and text line information to guide the model to focus on the global and local features of distorted paper, so as to reduce background interference and improve the readability of document images.
- We introduce cross-attention to explore more effective interaction between paper deformation trend and original distorted image.
- We conduct extensive experiments and show the state-of-the-art results on the existing prevalent benchmarks.

2. Related Work

2.1. Document Image Geometric Unwarping

The goal of rectifying distorted images is to enhance the visual quality of images, thereby reducing the difficulty of text extraction and improving the readability of paper. Early document image rectification methods performed parametric rectification by using cues easily observed from the image, such as text-lines proposed by Huang *et al.* [12], cylindrical surfaces [38, 6, 21, 18, 48, 27], document boundaries [3, 36] or laser beams from external devices used by Narain *et al.* [29].

In recent years, many works such as [26, 8, 1, 11, 14, 10] proposed to use deep learning to rectify distorted document images. These models solve the rectification problem by directly predicting the deformation field of the image.

Ma *et al.* [26] is the first work using deep neural networks on this task. They used a stacked UNet structure to predict the deformation field for each pixel in the warped document. This is an end-to-end structure that inverts the input image to obtain the result after the prediction result is obtained. In view of the lack of a large number of warped datasets to train the model, Das *et al.* [8] subsequently contributed a warped document image dataset (Doc3D) containing about 100,000 Distorted images. In addition, they

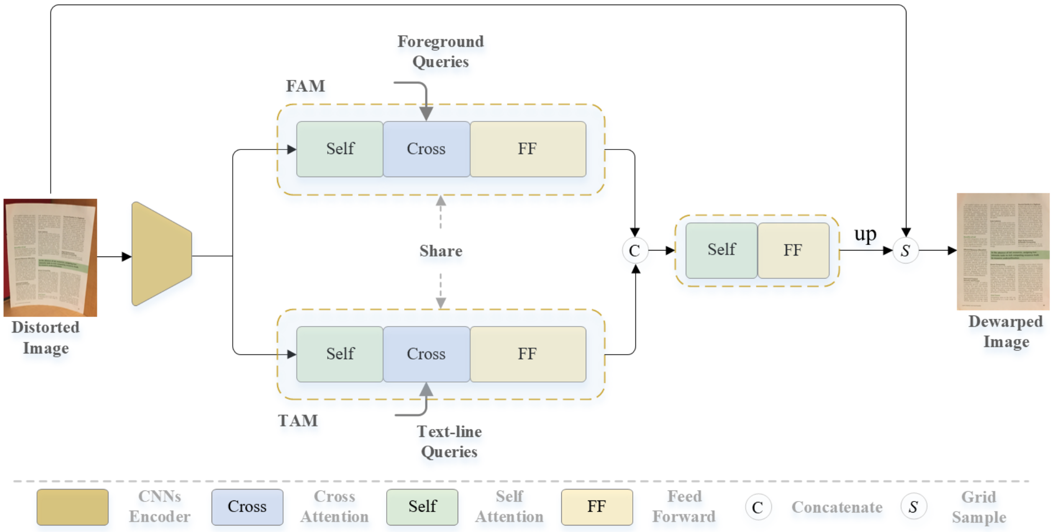

Figure 2. The overview architecture of our proposed method. FAM is **F**oreground **A**ttention **M**odule and TAM is **T**ext-lines **A**ttention **M**odule.

proposed a UNet structure network that uses 3D shape (World Coordinates) to replace the original warped document to predict the deformation field.

In order to reduce the interference of the background on the dewarping effect, Feng *et al.* [11] proposed to remove the background of the document before rectification, and introduced a transformer on this task for the first time. For reasons of improving the reading quality of the rectified document, other works of Jiang *et al.* [14] and Feng *et al.* [10] proposed to introduce the text-line feature in the document to enhance the rectification effect of text-line areas.

2.2. Attention in Vision

The attention mechanism focuses the attention of the network on the most important parts of the data by enhancing the weight of some parts of the input data of the neural network while weakening the weight of other parts. It was originally developed in 2017 for natural language processing tasks by Vaswani *et al.* [37], and greatly improved new state-of-the-art performance on many downstream tasks. However, in recent years, great success has been achieved in computer vision-related tasks such as image or video classification [39, 31, 2, 33, 7], image segmentation [49, 41, 4, 23], visual question answering [51, 34, 5, 46], and scene text recgonition [44].

Different from the self-attention mechanism, cross attention is able to exploit the intra-modal relations of each modality to complement each other and enhance the feature relations between different modalities. On multimodal tasks, cross attention is widely used to match the visual semantic similarity between images and texts [51, 34, 5, 46], and the key step is how to design fusion modules to effectively connect multimodal inputs. In this distorted image geometry rectification task, in order to minimize background interference, we propose to use the foreground information to aware corresponding regions in image features in a cross-attention manner. Furthermore, in order to increase the attention on the text-line regions, similar to the foreground cross-attention, we use the text-lines information to obtain the features of the text regions in the image, which can ensure that the same line of text is horizontal on the rectified image.

3. The Proposed Approach

In this section, we propose a noval framework for document image dewarping. The overview architecture of our model is shown in Figure 2.

Network Architecture. Our model consists of *CNNs Encoder*, the *Foreground and Text-line Attention Module,* and the *Transformer Decoder*. We use these three modules to obtain the features of foreground regions except background and text-line regions in distorted document images, respectively. After obtaining text-line and foreground information, cross-attention used to enhance the foreground and text-line attention. And then concatenate the feature map of these two cross feature map in the channel dimension, input it to the multi-layer decoder to predict the 2D grid coordinates, and finally use the coordinates to dewarp the input document image.

Given a warped document image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, we first resize it to $\mathbf{I}_r \in \mathbb{R}^{H_r \times W_r \times C_r}$, where $H_r = W_r = 224$ and $C_r = 3$ is the number of RGB channels of the original image. Then, $\mathbf{I}_r$ is fed into an encoder consisting of multiple layers of convolutional neural networks (CNNs) to obtain the feature map $\mathbf{F}_i \in \mathbb{R}^{H_i \times W_i \times C_i}$ of the original warped image. Then, through the cross-attention of fore-

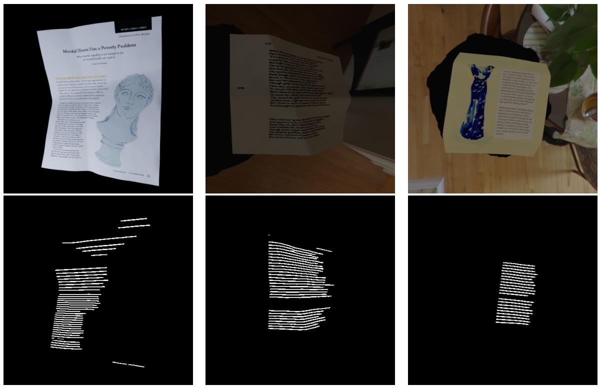

Figure 3. Text-line detection results of the Doc3D dataset. Row 1: Input warped images. Row 2: Detection results for text lines.

ground and text-line, the region of interest is extracted from the image features $\mathbf{F}_i$. Finally, a 2D dense grid coordinate $\hat{G} \in \mathbb{R}^{H_r \times W_r \times 2}$ (backward mapping or deformation field) is predicted through a transformer network composed of a multi-layer self-attention and feed-forward network architecture. According to the obtained grid coordinates, the input image can be dewarped to obtain an dewarped picture $\mathbf{D} \in \mathbb{R}^{H \times W \times 3}$. Each coordinate (x, y) in grid coordinates represents the position of the pixel value on the input image in the rectified image.

CNNs Encoder. For the resized image $\mathbf{I}_r$, we first use the convolutional neural network module as an encoder to extract image features. The encoder contains 3 layers of residual blocks, and each layer downsamples the feature map to half the size. The resolution of the final output feature map $\mathbf{F}_i \in \mathbb{R}^{H_i \times W_i \times C_i}$ is one-eighth of that of $\mathbf{I}_r$, Where $H_i = H_r/8$, $W_i = W_r/8$, C_i is the number of channels of the feature map $\mathbf{F}_i$. We set $\mathbf{C}_i$ to 256.

Foreground and Text-line Queries. In the previous work on distorted image rectification, there have been related works on foreground [11, 10] and text-line mask [14] extraction. We initialize the foreground and text-line queries on the basis of these two works. For foreground mask extraction, Feng *et al.* [11] used a lightweight semantic segmentation architecture network to predict the confidence map of foreground. The network is an Encoder-Decoder structure. We do not directly use the resulting binarized mask. Instead, use the six-layer feature maps in the decoder, and first resize them to the same size as $\mathbf{F}_i$, which is one-eighth the size of $\mathbf{I}_r$. And after concatenating in the channel dimension, use a layer of 3×3 convolution module and BatchNorm [13] to reduce the number of channels to 256. Compared with directly using the binary mask output by the last layer of the network, this method can obtain richer image feature information.

For the text-lines information, since the Doc3D training dataset proposed by DewarpNet [8] used by our model lacks corresponding text-line annotations, we use the detection text-line method proposed in RDGR [14] to obtain text-line features. This detection method used the U-Net network to train on PubLayNet [50] and the Scanned Script Dataset[1]. Its detection effect on the Doc3D dataset is shown in Figure 3. Although the text-line detected by this method has certain errors due to the deformation of the image, illumination effects and other reasons, it is within the network fault tolerance range. For text-line queries initialization, we use the last feature map of the U-Net network decoder (not the final confidence map). Its resolution is consistent with U-net network input. Then, we apply a layer of 3×3 convolution and BatchNorm [13] to change the number of channels to 256, and finally bilinear interpolation is utilized to downsample the feature map size to one-eighth the size of $\mathbf{I}_r$.

Foreground and Text-line Attention Module. For the initialized foreground queries, we use the cross-attention mechanism to extract features in distorted images. This module consists of three parts: *self-attention*, *cross-attention*, and *feed forward networks*. First, the image feature map $\mathbf{F}_i$ is flattened, and then the long-distance dependencies between patches are calculated using self-attention.

However, in the process of flattening the feature map, 2D coordinate information will be lost, which is crucial to obtain the local and global relationship, so we add a learnable 2D coordinate information. The same operation is performed on foreground queries. And then, in order to pay more attention to the foreground area and ignore the background, we use a foreground-guided cross-attention module. Specifically, a multi-head cross-attention layer is applied after self-attention. We utilize image features $\mathbf{F}_i$ as key and value, and the foreground feature $\mathbf{F}_f$ as query. Similar to the foreground-query cross-attention mechanism, in order to extract the features of the text-line positions of interest. The text-line features $\mathbf{F}_t$ is treated as query. Finally, we utilize a layer of convolutions to enhance local feature extraction after obtaining the attention feature map. The formula for the foreground query and text-line query cross-attention mechanism is as follows:

$$\begin{aligned} CA_f = CA(Q,K,V) &= CA(F_f, F_i, F_i) \\ &= \textit{softmax}(\frac{F_f F_i^{\mathrm{T}}}{\sigma})F_i \end{aligned} \tag{1}$$

$$\begin{aligned} CA_t = CA(Q,K,V) &= CA(F_t, F_i, F_i) \\ &= \textit{softmax}(\frac{F_t F_i^{\mathrm{T}}}{\sigma})F_i \end{aligned} \tag{2}$$

where CA is Cross-Attention and σ is a scaling factor. CA_f and CA_t are the foreground and text-line cross-attention map respectively.

Transformer Decoder. After extracting the interesting feature $\mathbf{C}A_f$ and $\mathbf{C}A_t$ of the foreground and text-line, we concatenate the two features and feed them into the transformer

[1] `https://github.com/zzzDavid/ICDAR-2019-SROIE`

Method	Pub.	MS-SSIM ↑	LD ↓	AD ↓	ED ↓	CER ↓
Distorted	-	0.2459	20.51	1.0134	2111.56	0.5352
DocUNet [26]	CVPR'18	0.4103	14.19	-	1933.66	0.4632
DocProj [20]	TOG'19	0.2946	18.01	-	1712.48	0.4267
DewarpNet [8]	ICCV'19	0.4735	8.39	0.4260	885.90	0.2373
FCN-based [42]	DAS'20	0.4288	7.75	0.4017	1792.60	0.4213
PWUNet [9]	ICCV'21	0.4915	8.64	-	1069.28	0.2677
DocTr [11]	ACM MM'21	0.5105	7.76	**0.3682**	724.84	0.1832
DDCP [43]	ICDAR'21	0.4726	8.97	0.4287	1411.38	0.3573
FDRNet [45]	CVPR'22	**0.5420**	8.21	-	829.78	0.2068
RDGR [14]	CVPR'22	0.4950	8.51	0.4382	729.52	0.1717
DocGeoNet [10]	ECCV'22	0.5040	**7.71**	0.3800	713.94	0.1821
Ours	-	0.4978	8.43	0.3761	**697.52**	**0.1705**

Table 1. On the DocUNet Benchmark [26], compare our proposed with existing methods in terms of Multi-Scale Structural Similarity (MS-SSIM), Local Distortion (LD) and OCR accuracy (ED and CER). "↑" indicates the higher the better and "↓" means the opposite.

decoder composed of multi-head self-attention mechanism to predict the grid map. Consistent with the cross-attention module, we also add 2D learnable position coordinates. The structure of each decoder layer consists of *self-attention* and *feed-forward network*. Finally, we adopt the upsampling method in DocTr [11] and GeoDewarpNet [10] to obtain high-resolution 2D deformation field $\mathbf{G}$. This upsampling method trains a learnable weight on the basis of initializing the grid coordinates to perform weighted optimization on it.

Training Loss Function. The loss function of our distorted document image rectification model is to compute the L_1 distance between the predicted grid coordinate $\hat{\boldsymbol{G}}$ and the label $\mathbf{G}$. The formula as follows:

$$\mathcal{L}_{grid} = \left\| \mathbf{G} - \hat{\boldsymbol{G}} \right\|_{\mathbf{1}} \tag{3}$$

4. Experiments

In this section, we first review the existing training dataset (Doc3D) and evaluation benchmark (DocUNet and DIR300 dataset) for distorted image dewarping. Second, we introduce the evaluation metrics and training details of this method. And then, we evaluate our document image rectification method on two datasets, DocUNet Benchmark [26] and DIR300 [10]. Finally, we demonstrate our rectification performance on several evaluation metrics.

4.1. Dataset

Doc3D. We train the geometric rectification model on the Doc3D dataset contributed by DewarpNet [8]. Doc3D is the largest distortion dewarping dataset so far, containing a total of 100K distorted images. It is created by about 4000 real document images and rendering software. Among them, different camera positions and various illumination are applied when rendering. For each distorted document image, the corresponding labels include 3D coordinate maps, albedo maps, normals, depth maps, UV maps and backward mapping map.

DocUNet Benchmark. In current deep learning-based document dewarping methods, the DocUNet Benchmark [26] dataset is widely used. This evaluation dataset contains 130 distorted images in natural scenes captured by mobile devices. In addition, for better comparison with previous methods, we follow the proposal in DocGeoNet [10] and rotate the 127^{th} and 128^{th} images in DocUNet [26] that do not match the labels by 180 degrees.

DIR300 Dataset. DIR300 dataset is a new test dataset proposed by DocGeoNet [10] captured by moving camera. It contains 300 images of real distorted documents involving more complex backgrounds, distorted degrees, and various lighting conditions.

4.2. Evaluation metrics

MS-SSIM, LD and AD. We follow previous methods [26, 8, 1, 11, 14, 10] to perceive image quality using an image-based Multi-Scale Structural Similarity method (MS-SSIM) [40]. Computationally dense SIFT-flow [22], Local Distortion (LD) [47] computes the average of the L_2 distances between all pixels in the ground-truth scanned image and the rectified image pixels, which measures the average local deformation of the rectified image. Following DocUNet [26], all rectified images and their labels are resized to 598400 pixel area. We use the code provided by DocUNet [26] for evaluation. Aligned Distortion (AD) is a more robust evaluation metric proposed by [25], which aligns unwarped images and ground truth scans by unifying translation and scale before computing distortions.

CER and ED. We use Character Error Rate (CER) [28] and Edit Distance (ED) [19] to measure the performance of our method. ED is a string metric that measures the difference between two character sequences. The edit distance between two strings is the minimum number of single-character edits (insertions (i), deletions (d), or substitutions (s)) required to transform one string into the reference string. Then, the calculation of CER is: $(i+d+s)/N$, where N is the number of reference strings. Following the latest methods DocTr [11] and DocGeoNet [10], we use Tesser-

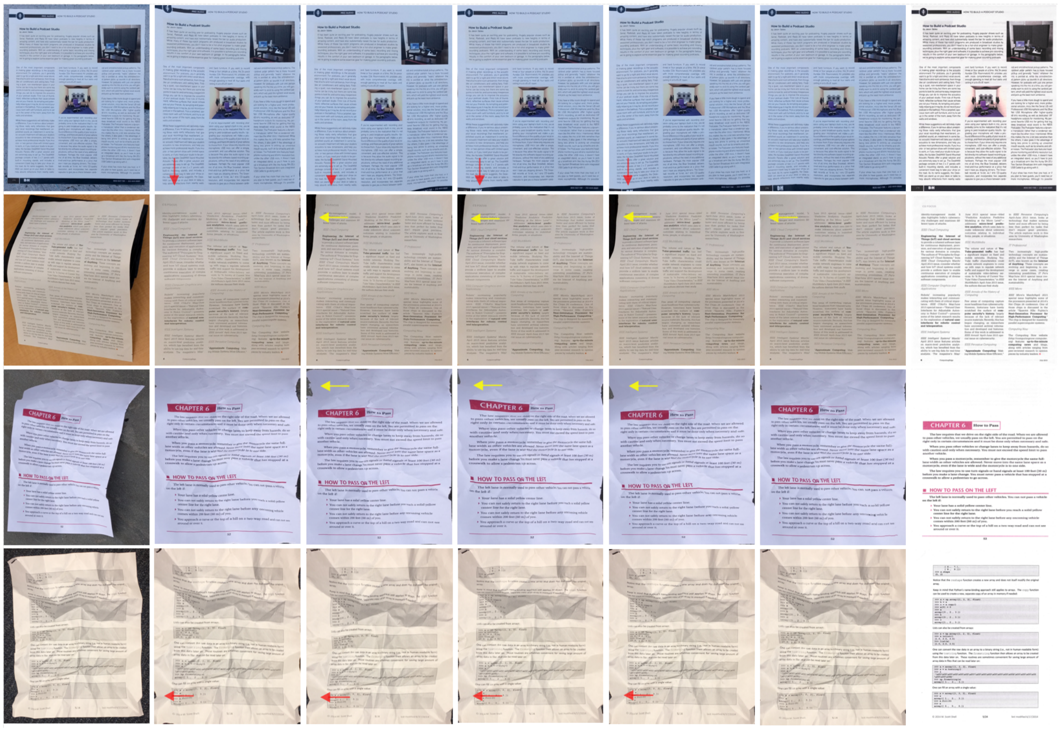

Figure 4. Comparisons on DocUNet Benchmark [26]. Column 1: Distorted images. Columns 2-6: Results of DewarpNet [8], DocTr [11], RDGR [14], DocGeoNet [10] and ours method. Last column: Flatbed scanned images. We highlight the differences by using red and yellow arrows.

Method	Pub.	MS-SSIM ↑	LD ↓	AD ↓	ED ↓	CER ↓
Distorted	-	0.3169	39.58	0.7707	1500.56	0.5234
DocProj [20]	TOG'19	0.3246	30.63	-	958.89	0.3540
DewarpNet [8]	ICCV'19	0.4921	13.94	0.3320	1059.57	0.3557
DocTr [11]	ACM MM'21	0.6160	7.21	0.2552	699.63	0.2237
DDCP [43]	ICDAR'21	0.5524	10.97	0.3554	2130.01	0.5524
DocGeoNet [10]	ECCV'22	**0.6380**	**6.40**	**0.2418**	664.96	0.2189
Ours	-	0.6070	7.68	0.2440	**652.80**	**0.2115**

Table 2. On the DIR300 Dataset [10], compare our proposed with existing methods.

act (v5.0.1) [32] as an OCR engine to recognize text in the images. And, following the settings in DocUNet [26], DewarpNet [8] and RDGR [14], 50 images are selected from DocUNet Benchmark [26] to evaluate OCR performance. On DIR300 dataset [10], we compute CER and ED for 90 text-rich images following the setting in DocGeoNet [10].

4.3. Implementation Details

We implement the entire model architecture on the PyTorch [30] framework. We train our document rectification model on the Doc3D dataset of 100,000 images. The image size for training is 224 × 224. We set the layers of cross-attention and transformer decoder to 12 and 6, respectively. We use the AdamW optimizer [24] with a batch size of 24. The initial learning rate is 1×10^{-4}. The whole model is trained on two NVIDIA A100 GPUs for 30 epochs until the model converges.

4.4. Experimental Results

We use OCR performance (CER and ED) and image distortion metrics (MS-SSIM and LD) to compare our method with existing deep learning-based state-of-the-art rectification models on DocUNet Benchmark [26] and DIR300 Dataset [10].

Evaluation on DocUNet Benchmark. As shown in Table 1, in the setting of 50 images, compared with the existing works, our method for distorted document image dewarping improves both ED and CER. Among them, the ED metric dropped below 700 for the first time to reach 697.52, an increase of 2.3%. At the same time, it reached 0.1705 on the CER. Our method outperforms almost all non-textline methods [26, 20, 8, 42, 9, 11, 43, 45] on distortion metrics. It can be seen that making the model focus on the features of the text-line regions significantly improves the rectification effect. Overall, in addition, compared with RDGR [14] and DocGeoNet [10], which also use text-line for distor-

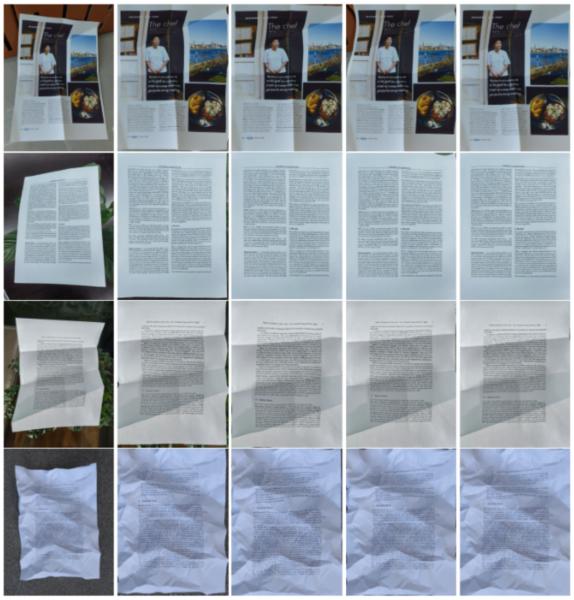

Figure 5. Comparisons on DIR300 [10]. Column 1: Distorted images. Column 2-4: Results of DewarpNet [8], DDCP [43], DocGeoNet [10]. Column 5: Rectification results of our model.

tion images dewarping, our method performs better in OCR evaluation metrics.

Evaluation on DIR300 Dataset. In the evaluation of the DIR300 dataset, the results are shown in Table 2. We follow the settings in DocGeoNet [10], and select 90 of the pictures for evaluation in terms of OCR accuracy. For previous state-of-the-art methods, our method has improved both CER and ED, and ED has increased by 1.8% and dropped to 652.8. Compared with the existing state-of-the-art methods, from the experimental results, our method can more effectively improve the OCR performance of the rectified image. In order to show the rectification effect more intuitively, we visualize and compare the results between existing and our methods on DocUNet Benchmark [26] and DIR300 Dataset [10], as shown in Figure 4 and Figure 5. It can be seen from the first two rows of these two figures that our method pays more attention to the foreground of distorted document images. The last two lines of images show that the dewarping effect of the text-line is more horizontal after the introduction of the text-line information. This is crucial for improving OCR accuracy and paper readability.

4.5. Experimental Analysis

Visualizations of OCR Results. We visualized the recognition results of OCR after image rectification, as shown in the Figure 6. We compared the effect of the distorted images, dewarping results of DocGeoNet [10] and our method on OCR recognition. We highlight the text area detected by Tesseract (v5.0.1) [32] OCR engine with three different colors. It can be seen from the figure that the OCR performance can be greatly improved after dewarping the distorted image. By fusing the information of text-lines, the model can be made to pay more attention to the text area, improving the reading quality of the document image.

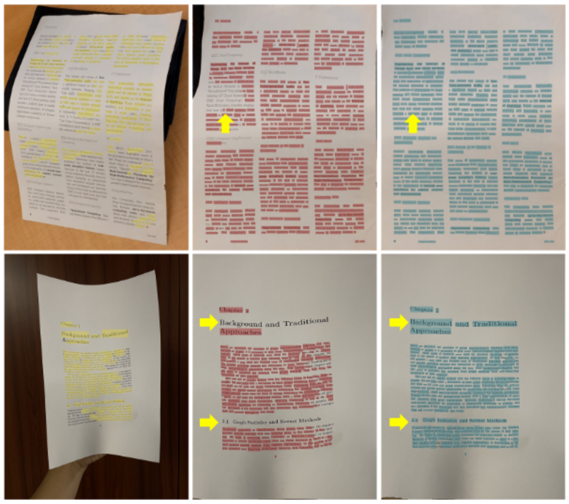

Figure 6. **Visualizations of OCR results.** Left to right: Original deformed images, rectified by DocGeoNet [10], rectified by ours method. The color is the highlighted text detection results. Yellow arrows indicate differences.

Visualizations of Cross-attention Maps. In order to further show the feature extraction effect of foreground and text-lines, we provide the attention map visualization for the foreground and text-line attention map from our cross-attention module (FAM and TAM) in Figure 7. We can see that our model can acquire the ability to localize the foreground and text-line regions.

Bad Case Analysis. Although our method improves the OCR accuracy to some extent, it has not reached SOTA on MS-SSIM, LD and AD. According to our analysis, as shown in the Figure 8, after observing the rectified samples, we found that the characteristics of the input distorted image itself, such as the influence of illumination condition, shadows, and background that is very close to the foreground boundary, although these do not affect the attention feature extraction of text-line regions. It is also because of these attributes that the effect of our method to segment the margins between foreground and background on similar images with these characteristics still needs to be improved.

Besides, We analyzed the influence of the text content ratio on the DocUNet Benchmark. Sorting in ascending order of the proportion of text content, the first 30% of images as **Low**, the next 40% of images as **Middle**, and the last 30% of images as **High**. As shown in table 3, Due to MS-SSIM is sensitive to pixel location, highly subtle image differences can result in large MS-SSIM differences. Since the connectivity of text area is lower than general images, the more text content may cause lower MS-SSIM. This phenomenon is also manifested in AD. As the ratio of text increases, LD tends to be better, which indicates that the text line feature

Text Content Ratio	MS-SSIM ↑	LD ↓	AD ↓
Low	0.54	10.74	0.35
Middle	0.51	8.27	0.35
High	0.44	6.34	0.44

Table 3. The relationship between the proportion of text content and each evaluation metric.

FAM	TAM	MS-SSIM ↑	LD ↓	AD ↓	ED ↓	CER ↓
✓		0.4785	8.53	0.3619	761.22	0.1952
	✓	0.4778	8.86	**0.3529**	748.46	0.1889
✓	✓	**0.4978**	**8.43**	0.3761	**697.52**	**0.1705**

Table 4. Ablation experiments. FAM is **F**oreground **A**ttention **M**odule and TAM is **T**ext-lines **A**ttention **M**odule.

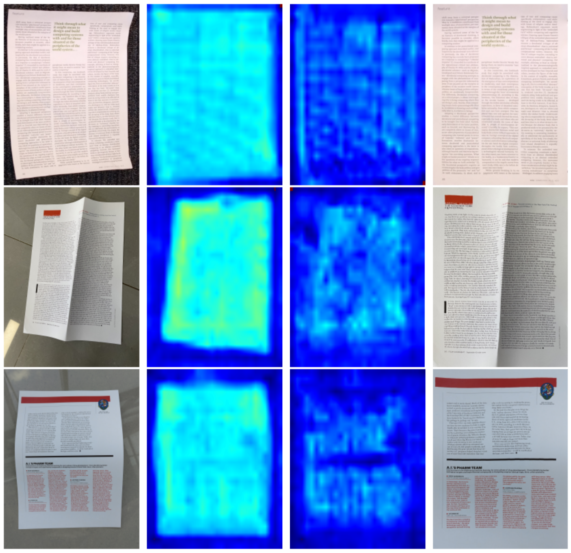

Figure 7. **Visualizations of attention maps.** Left to right: distorted images, foreground attenion maps, text-line attention maps, rectification results of our model.

is beneficial for local distortion correction, which is also reflected from the CER and ED metrics.

Ablation Studies. Table 4 shows the performance of our method when foreground and text-line information are used respectively. We compared the performance when only one of the two features mentioned above is used. This experimental results demonstrate that using foreground and text-line feature extraction modules together, the MS-SSIM and Local Distortion (LD) are increased by at least 4% and 1.2% respectively. It can be seen that the overall rectification effect of the distorted document image is the best when the two features are combined to extract the region of interest in the original image. It is precisely due to the foreground contains the overall distorted outline of the deformed image, and the text-line contains the local deformation trend of the text area, so that the global and local information can be complementary with each other, and the overall rectification effect can be greatly improved.

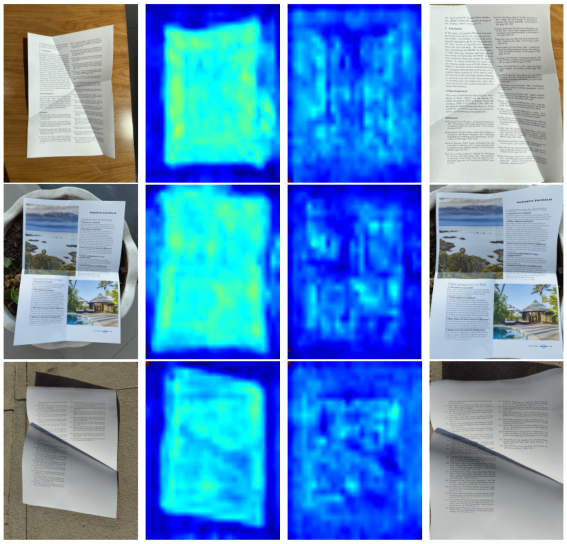

Figure 8. **Visualizations of attention maps for bad case analysis.** Left to right: distorted images, foreground attention maps, text-line attention maps, rectification results of our model. Row 1: influence of illumination condition. Row 2: the boundary of the foreground close to background. Row 3: the shadow interference of the document image.

5. Conclusion

In this work, we explored how to optimize the rectification of distorted document images from the perspective of readability. We propose an efficient method for warped document rectification by focusing on the foreground and text-line regions of the original deformed image separately via cross-attention. We use this method to improve the readability of the rectified document images from the perspective of global and local features. The effectiveness of this feature fusion mechanism is proved by experiments on two public benchmark datasets. In the future, since the current training dataset lacks text-line annotations, we will further explore how to obtain more accurate text-line annotations. This is also a limitation of our method. Besides, for extracting distorted image features, we will also explore more fine-grained and multi-attribute feature fusion methods to improve image quality and OCR performance.

Acknowledgement: We thank the reviewers for their thoughtful and constructive comments. This work was supported by the National Key R&D Program of China (2022ZD0116002), Natural Science Foundation of China (62276075, 61872113), Science and Technology Planning Project of Shenzhen (JCYJ20190806112210067), Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies, China (2022B1212010005) and China Mobile Information Technology Co.,Ltd. It was also supported by Music Style Transfer for Anxiety Therapy (P0042962), Artificial Intelligence and Robotics (AIR) Group: Artifical Intelligence Art (P0033738).

References

[1] Hmrishav Bandyopadhyay, Tanmoy Dasgupta, N. Das, and Mita Nasipuri. A gated and bifurcated stacked u-net module for document image dewarping. *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10548–10554, 2020.

[2] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V. Le. Attention augmented convolutional networks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3285–3294, 2019.

[3] M. S. Brown and Yau-Chat Tsoi. Geometric and shading correction for images of printed materials using boundary. *IEEE Transactions on Image Processing*, 15:1544–1554, 2006.

[4] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G. Schwing. Mask2former for video instance segmentation. *ArXiv*, abs/2112.10764, 2021.

[5] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. *ArXiv*, abs/2102.02779, 2021.

[6] Frédéric Courteille, Alain Crouzil, Jean-Denis Durou, and Pierre Gurdjos. Shape from shading for the digitization of curved documents. *Machine Vision and Applications*, 18:301–316, 2007.

[7] Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *ArXiv*, abs/2106.04803, 2021.

[8] Sagnik Das, Ke Ma, Zhixin Shu, Dimitris Samaras, and Roy Shilkrot. Dewarpnet: Single-image document unwarping with stacked 3d and 2d regression networks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 131–140, 2019.

[9] Sagnik Das, Kunwar Yashraj Singh, Jon Wu, Erhan Bas, Vijay Mahadevan, Rahul Bhotika, and Dimitris Samaras. End-to-end piece-wise unwarping of document images. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4248–4257, 2021.

[10] Hao Feng, Wen gang Zhou, Jiajun Deng, Yuechen Wang, and Houqiang Li. Geometric representation learning for document image rectification. *ArXiv*, abs/2210.08161, 2022.

[11] Hao Feng, Yuechen Wang, Wen gang Zhou, Jiajun Deng, and Houqiang Li. Doctr: Document image transformer for geometric unwarping and illumination correction. *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

[12] Zuming Huang, Jie Gu, Gaofeng Meng, and Chunhong Pan. Text line extraction of curved document images using hybrid metric. *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 251–255, 2015.

[13] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ArXiv*, abs/1502.03167, 2015.

[14] Xiangwei Jiang, Rujiao Long, Nan Xue, Zhibo Yang, Cong Yao, and Guisong Xia. Revisiting document image dewarping by grid regularization. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4533–4542, 2022.

[15] Tae Ho Kil, Wonkyo Seo, Hyung Il Koo, and Nam Ik Cho. Robust document image dewarping method using text-lines and line segments. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 01:865–870, 2017.

[16] Beom Su Kim, Hyung Il Koo, and Nam Ik Cho. Document dewarping via text-line based optimization. *Pattern Recognit.*, 48:3600–3614, 2015.

[17] Hyung Il Koo and Nam Ik Cho. Rectification of figures and photos in document images using bounding box interface. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3121–3128, 2010.

[18] Hyung Il Koo, Jinho Kim, and Nam Ik Cho. Composition of a dewarped and enhanced document image from two view images. *IEEE Transactions on Image Processing*, 18:1551–1562, 2009.

[19] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710, 1965.

[20] Xiaoyu Li, Bo Zhang, Jing Liao, and Pedro V. Sander. Document rectification and illumination correction using a patch-based cnn. *ACM Transactions on Graphics (TOG)*, 38:1 – 11, 2019.

[21] Jian Liang, Daniel DeMenthon, and David S. Doermann. Geometric rectification of camera-captured document images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:591–605, 2008.

[22] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:978–994, 2011.

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021.

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.

[25] Ke Ma, Sagnik Das, Zhixin Shu, and Dimitris Samaras. Learning from documents in the wild to improve document unwarping. *ACM SIGGRAPH 2022 Conference Proceedings*, 2022.

[26] Ke Ma, Zhixin Shu, Xue Bai, Jue Wang, and Dimitris Samaras. Docunet: Document image unwarping via a stacked u-net. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4700–4709, 2018.

[27] Gaofeng Meng, Yuanqi Su, Ying Wu, Shiming Xiang, and Chunhong Pan. Exploiting vector fields for geometric rectification of distorted document images. In *European Conference on Computer Vision*, 2018.

[28] Andrew C. Morris, Viktoria Maier, and Phil D. Green. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Interspeech*, 2004.

[29] Rahul Narain, Tobias Pfaff, and James F. O'Brien. Folding and crumpling adaptive sheets. *ACM Transactions on Graphics (TOG)*, 32:1 – 8, 2013.

[30] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zach DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[31] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *IEEE Workshop/Winter Conference on Applications of Computer Vision*, 2018.

[32] R. Smith. An overview of the tesseract ocr engine. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, 2:629–633, 2007.

[33] A. Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, P. Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16514–16524, 2021.

[34] Wei Suo, Mengyang Sun, Peng Wang, and Qi Wu. Proposal-free one-stage referring expression via grid-word cross-attention. In *International Joint Conference on Artificial Intelligence*, 2021.

[35] Yuandong Tian and Srinivasa G. Narasimhan. Rectification and 3d reconstruction of curved document images. *CVPR 2011*, pages 377–384, 2011.

[36] Yau-Chat Tsoi and M. S. Brown. Multi-view document rectification using boundary. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[37] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017.

[38] Toshikazu Wada, Hiroyuki Ukida, and Takashi Matsuyama. Shape from shading with interreflections under a proximal light source: Distortion-free copying of an unfolded book. *International Journal of Computer Vision*, 24:125–135, 1997.

[39] X. Wang, Ross B. Girshick, Abhinav Kumar Gupta, and Kaiming He. Non-local neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2017.

[40] Zhou Wang, Alan Conrad Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004.

[41] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, José Manuel Álvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Neural Information Processing Systems*, 2021.

[42] Guo-Wang Xie, Fei Yin, Xu-Yao Zhang, and Cheng-Lin Liu. Dewarping document image by displacement flow estimation with fully convolutional network. *ArXiv*, abs/2104.06815, 2020.

[43] Guo-Wang Xie, Fei Yin, Xu-Yao Zhang, and Cheng-Lin Liu. Document dewarping with control points. In *IEEE International Conference on Document Analysis and Recognition*, 2022.

[44] Xudong Xie, Ling Fu, Zhifei Zhang, Zhaowen Wang, and Xiang Bai. Toward understanding wordart: Corner-guided transformer for scene text recognition. In *European Conference on Computer Vision*, 2022.

[45] Chuhui Xue, Zichen Tian, Fangneng Zhan, Shijian Lu, and Song Bai. Fourier document restoration for robust document dewarping and recognition. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4563–4572, 2022.

[46] Ziyi Yang, Yuwei Fang, Chenguang Zhu, Reid Pryzant, Dongdong Chen, Yu Shi, Yichong Xu, Yao Qian, Mei Gao, Yi-Ling Chen, Liyang Lu, Yujia Xie, Robert Gmyr, Noel C. F. Codella, Naoyuki Kanda, Bin Xiao, Yuanxun Lu, Takuya Yoshioka, Michael Zeng, and Xuedong Huang. i-code: An integrative and composable multimodal learning framework. *ArXiv*, abs/2205.01818, 2022.

[47] Shaodi You, Yasuyuki Matsushita, Sudipta Sinha, Yu-Ki Bou, and Katsushi Ikeuchi. Multiview rectification of folded documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:505–511, 2016.

[48] Li Zhang, Andy M. Yip, M. S. Brown, and Chew Lim Tan. A unified framework for document restoration using inpainting and shape-from-shading. *Pattern Recognit.*, 42:2961–2978, 2009.

[49] Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021.

[50] Xu Zhong, Jianbin Tang, and Antonio Jimeno-Yepes. Publaynet: Largest dataset ever for document layout analysis. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022, 2019.

[51] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8743–8752, 2020.