DocReal: Robust Document Dewarping of Real-Life Images via Attention-Enhanced Control Point Prediction

Fangchen Yu[1]*, Yina Xie[2], Lei Wu[2], Yafei Wen[2], Guozhi Wang[2]
Shuai Ren[2], Xiaoxin Chen[2], Jianfeng Mao[1,3], Wenye Li[1,3]†
[1]The Chinese University of Hong Kong, Shenzhen, China
[2]vivo AI Lab, Shenzhen, China
[3]Shenzhen Research Institute of Big Data, Shenzhen, China

Abstract

Document image dewarping is a crucial task in computer vision with numerous practical applications. The control point method, as a popular image dewarping approach, has attracted attention due to its simplicity and efficiency. However, inaccurate control point prediction due to varying background noises and deformation types can result in unsatisfactory performance. To address these issues, we propose a robust document dewarping approach for real-life images, namely DocReal, which utilizes Enet to effectively remove background noise and an attention-enhanced control point (AECP) module to better capture local deformations. Moreover, we augment the training data by synthesizing 2D images with 3D deformations and additional deformation types. Our proposed method achieves state-of-the-art performance on the DocUNet benchmark and a newly proposed benchmark of 200 Chinese distorted images, exhibiting superior dewarping accuracy, OCR performance, and robustness to various types of image distortion.

1. Introduction

The rising trend of using smartphones to digitize documents over conventional scanners has intensified the challenges of document image dewarping. Captured images frequently exhibit diverse physical deformations, shooting angles, and ambient disturbances, resulting in a wide range of distorted document images. As a result, document image dewarping has emerged as a fervent research focus recently [5, 8, 21, 35].

Traditional image dewarping techniques hinge on the 3D structures of distorted images, segmented into three primary categories. The first category emphasizes 3D image structure reconstruction, either through auxiliary tools like structured beams and lasers [2, 3, 25, 38] or through multi-view images [17, 28, 29, 36]. However, incorporating extra devices limits their broad applicability. The secondary approach leverages parameterized modeling to fit the 3D structure [4, 12, 14, 15], involving optimization under constraints. Nevertheless, their oversimplifications fall short of accurately capturing real deformations, leading to unsatisfactory dewarping performance. The last strategy uses 3D structures as supervised signals to train a pixel-wise learning model, employing 3D coordinate maps [5, 9, 22] or 3D shape patches [6]. While the 3D structure aids enhance deformation predictions, the overheads in data preparation using depth cameras and rendering software, and model training remain considerable.

Lately, deep learning models leveraging 2D images have gained traction due to the simplicity of obtaining 2D training datasets. Many utilize an end-to-end framework for pixel-wise prediction, employing stacked UNet [21], FCN [34], Transformer [7, 8], or other networks [20]. Despite the partial success brought by pixel-wise regression, the DDCP model [35], based on control points, offers a flexible alternative using a lightweight network. Instead of traditional backward mapping, this model rectifies images by mapping control points from the distorted image to reference points on the flattened version, as illustrated in Fig. 1.

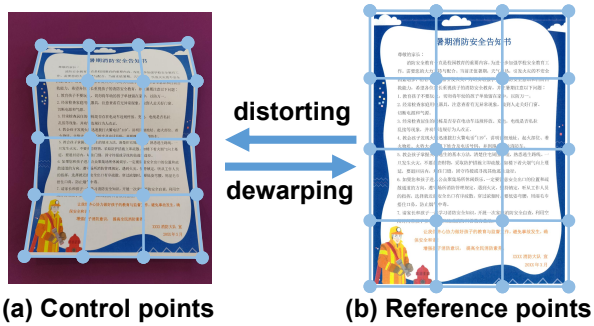


Figure 1. **Diagram of the control point based image dewarping.**

*Work done during Fangchen Yu's internship at vivo AI Lab
†Corresponding author: wyli@cuhk.edu.cn

The DDCP model [35], while straightforward in deployment, grapples with three major challenges: **1)** residual background in the rectified images, **2)** reduced text readability, and **3)** a notable disparity between training and real-world data. Various background disturbances and deformations can hinder precise control point prediction, complicating accurate local deformation capture on documents. Moreover, 2D synthetic training datasets produced by functions typically deviate from real-world data, constraining the model's efficacy.

To address these issues, we propose a robust control point based model and augment training data synthesis with the following contributions.

- **More robust model:** We propose a robust document dewarping framework for real-life distorted images, namely DocReal, which adopts Enet [20] to effectively remove background noise and an attention-enhanced control point (AECP) module to better capture local deformations. This approach achieves improved readability and robustness of dewarping images under different environmental conditions and deformation types with state-of-the-art performance on two benchmarks.
- **More realistic training data:** We utilize 3D images in the Doc3D [5] dataset to synthesize 2D images with 3D deformations. Furthermore, we augment the training data by adding various noises and randomly selected backgrounds. We also simulate additional four types of curling and folded deformations by formulas to enrich the deformation types of training data.
- **More comprehensive benchmark:** We provide the first Chinese document image benchmark of 200 distorted images that are more representative of real-life scenarios, including work, study, and daily life scenarios. This benchmark consists of 50 document contents and each with four types of deformation, facilitating evaluations of document dewarping in real-life.

2. Related Work

In recent years, 2D image based deep learning techniques have emerged as a promising alternative to traditional strategies for document image dewarping, which can be broadly categorized into two types. The first type involves pixel-wise flow regression, which predicts a pixel-wise displacement field or backward mapping to rectify the distorted document image. Representative models in the realm of pixel-wise flow regression encompass DocUNet [21], DocProj [19], FCN-based method [34], and DocTr [8]. Each of these employs distinct network architectures to enhance document dewarping efficacy. For instance, DocTr [8] uses the Transformer [30] to improve the backward mapping field predictions. Furthermore, DocTr++ [7] augments the capabilities of DocTr [8] with a hierarchical network structure, adept at processing a range of unrestricted document images.

Beyond pixel-wise flow regression, numerous other 2D image based deep learning techniques have been introduced for document image dewarping. Methods based on control points, like DDCP [35] and CGU-Net [31], estimate control points on a 2D dewarping grid to rectify distorted images using interpolation techniques [24]. Marior [37] first predicts key points to detect the document body, followed by iterative rectification to obtain the displacement flow. RDGR [13] recognizes image boundaries and textlines via UNets [27] and introduces a grid regularization to calculate a uniform forward map grounded on constraints. PaperEdge [20] engages Enet and Tnet to discern image contours and learn local deformation, respectively. Each of these methods has proposed different enhancements to improve document dewarping performance.

3. Methodology

Despite achieving impressive results, the control point based DDCP method [35] still faces significant limitations, particularly when it comes to accurately placing control points on text in different shooting environments, text types, and noisy backgrounds. Poor handling of this task can result in severe text deformation and background residue. To overcome these challenges, we draw inspiration from the PaperEdge method [20] and develop a new pipeline (see Fig. 2), namely DocReal, which first uses Enet to detect document edge information, remove background noise, and extract the document's main body while extending it globally. We then utilize an attention-enhanced control point (AECP) network to better capture local deformations of the document, predicting more accurate control points compared to the DDCP method [35]. Our method is a robust control point framework that significantly removes background noise and enhances readability under varying environmental conditions and text types.

3.1. Network Architecture

Our pipeline comprises two sub-networks: Enet and AECP. Enet is a fully-convolutional encoder-decoder that uses 6 residual blocks [11] in the encoder and 4 residual blocks in the decoder, as proposed by [20]. Enet is trained on synthetic images and weakly supervised on real images, producing a coarse result with a global shape through edge-based dewarping (for more details, see [20]).

On the other hand, AECP is an individual network that plays a crucial role in our document dewarping pipeline. It includes four sub-modules that work together to improve control point prediction accuracy for local deformations. The first sub-module (green in Fig. 2) extracts shallow features of the input image $I \in \mathbb{R}^{992\times992\times3}$ using two convo-

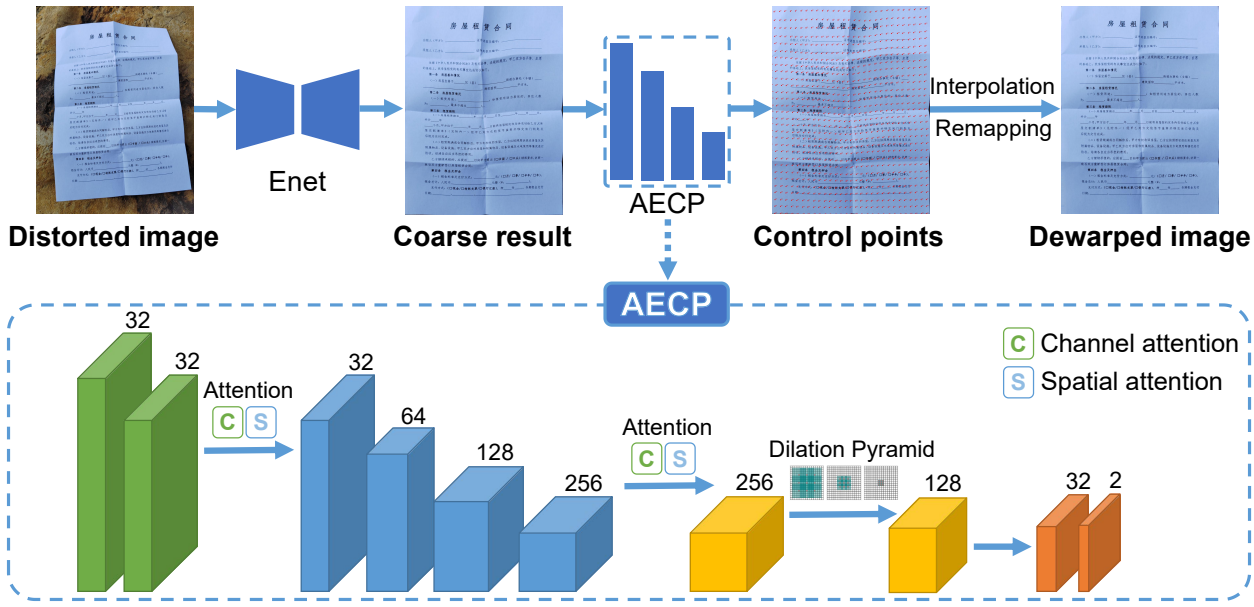


Figure 2. **Pipeline of DocReal**, a robust document dewarping approach of real-life images using attention-enhanced control point (AECP). Our pipeline consists of two sub-networks: Enet and AECP. Enet [20] performs an edge-based dewarping of the document, rectifying the overall shape to produce a coarse result. AECP then enhances this result by predicting accurate control points for local deformations. The final result is achieved through linear interpolation [1] and remapping from the coarse result to the dewarped image.

lutional layers with a stride of 2 and a 3×3 kernel size. The second sub-module (blue in Fig. 2) consists of four layers with 3×3 kernels for further deep feature extraction. Notably, both sub-modules are enhanced by an attention module [33] that leverages channel attention and spatial attention to prioritize critical information, such as shallow features (including light, shadow, and texture) and deep features (like table lines, text lines, and overall deformation).

The third sub-module (yellow in Fig. 2) utilizes a dilation pyramid [23] to broaden the perception of global deformation caused by the conduction of paper deformation. The dilation pyramid comprises six layers with a maximum dilation rate of 18. We concatenate the features from six layers, which have different scales, and then feed them into a 1×1 convolutional layer to obtain the global deformation feature. The fourth sub-module (orange in Fig. 2) uses a two-layer convolutional network to predict control points $P \in \mathbb{R}^{2\times31\times31}$. Each control point on the 31×31 grid has (x, y) coordinates that span the entire document, resulting in a robust control point framework that enhances readability under varying environmental conditions and text types.

Benefiting from these modules, the pipeline significantly removes background noises and enhances the readability of dewarped images, providing a robust and reliable solution.

3.2. Loss Function

Consistent with the DDCP method [35], we utilize a supervised manner to train the model with three parts of loss functions. The first part is the Smooth L_1-based localization loss [10, 26], which is used for regression on control points to predict their positions and is defined as follows.

$$L_{\text{loc}} = \frac{1}{N_c} \sum_{i=1}^{N_c} \text{smooth}_{L_1}(\hat{x}_i - x_i), \tag{1}$$

where $\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases}$, N_c is the number of control points, $\hat{x}_i$ and x_i denote the predicted 2D position of the i-th control point and the ground-truth, respectively. The second part models the neighboring relationship of control points with the L_{neigh} function, i.e.,

$$L_{\text{neigh}} = \frac{1}{N_c} \sum_{i=1}^{N_c} (\hat{\delta}_i - \delta_i)^2, \tag{2}$$

where δ denotes the local correlation between the control point and its vertical and horizontal neighbors. The third part uses L_1 loss for interval regression on reference points:

$$L_{\text{reg}} = \frac{1}{N_r} \sum_{j=1}^{N_r} \|\hat{d}_j - d_j\|_1, \tag{3}$$

where d denotes the interval between two horizontal or vertical points and $N_r = 2$. The total loss is a linear combination of these losses, with weights α and β (defaulted to 0.1 and 0.01, respectively), and more details see [35]:

$$L = L_{\text{loc}} + \alpha L_{\text{neigh}} + \beta L_{\text{reg}}. \tag{4}$$

3.3. Training Data Synthesis with 3D Deformation

Effective document image dewarping requires training data with authentic and diverse deformations. However, the training data for the DDCP method [35] is synthesized using two functions [4], resulting in a 2D mesh with significant differences from real-world deformations. While the Doc3D dataset [5] provides rich and realistic 3D deformations for training data, it cannot be directly used for training a control point network that requires 2D data.

To overcome this limitation, we propose a new method that synthesizes 2D training data with 3D deformation by the following four steps as shown in Fig. 3. First, we extract 3D mesh from a 3D image in the Doc3D dataset by sampling 31×31 control points from a 3D coordinate map, generating a point cloud from their 3D coordinates, and transforming it into a 3D mesh. Second, we randomly set camera distances and shooting angles in the 3D coordinate system to map control points in the 3D mesh to 2D control points. Third, we generate a 2D distorted image with 3D deformation by mapping the reference points of a scanned image onto the 2D control points and interpolating the pixels. Finally, we randomly add various noises to scanned images, such as moire patterns, fingerprints, shadows, and others, to enrich the synthetic images with random backgrounds.

The proposed data synthesis method can generate a significant amount of 2D training data with realistic and diverse 3D deformations, which is essential for training a control point network for document image dewarping.

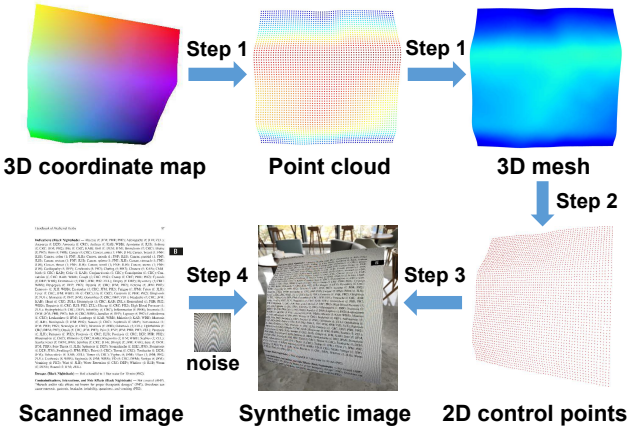


Figure 3. **Synthesis process of 2D data with 3D deformation.**

3.4. Training Data Augmentation

While the Doc3D dataset offers a range of deformations, it lacks the true curling of book types. To address this, we simulate the curling of books using formulas and add additional folded deformations for training data augmentation, thus achieving a more robust model performance.

As shown in Fig. 4, for curling types, **(a) unidirectional curling** refers to curling in the same direction outside two given key lines using the following function:

$$z_i = \cos(d_i \cdot v) - 0.5, \tag{5}$$

where only the z-axis values are modified, 0.5 is the original z value, d_i is the distance between the control point to the nearest line, and v is a hyper-parameter to control the degree of curling; **(b) S-shaped curling** refers to the bidirectional bending on both sides of the given key lines by:

$$z_i = \begin{cases} 0.5 + \cos(\frac{\pi}{2} - d_i \cdot v), & \text{if } i \in \text{region (1)}, \\ 0.5 + \cos(\frac{\pi}{2} + d_i \cdot v), & \text{if } i \in \text{region (2)}. \end{cases} \tag{6}$$

For folded types, **(c) corner folded** and **(d) edge folded** both refer to folding the control points outside the key lines inward or outward. Taking inward folding as an example:

$$z_i = 0.5 - d_i \cdot v. \tag{7}$$

By incorporating these additional deformation types into the training data set, we are able to improve the robustness of our dewarping model, as the model is exposed to a wider range of variations and deformations.

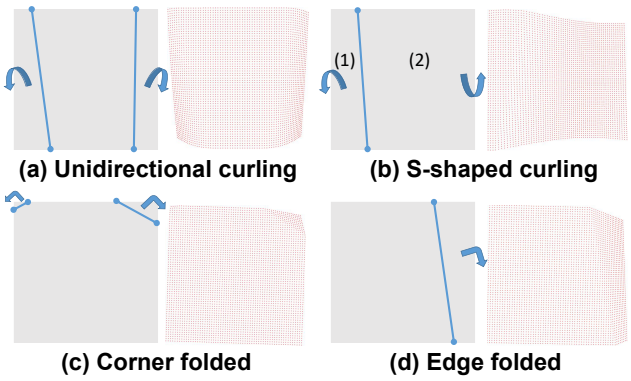


Figure 4. **Synthetic 3D deformation types.**

3.5. Implementation Details

We synthesized 300K 2D distorted images with 3D deformations and additional deformation types as the training dataset. We trained the model on a single NVIDIA Tesla V100 GPU for 40 epochs with a batch size of 16 and the Adam optimizer [16]. Due to commercial restrictions, we are unable to provide the source code currently. Instead, we provide rectified results and a newly proposed benchmark dataset[1] (which is introduced in Sec. 4.2), allowing for further exploration by the research community.

[1] https://github.com/irisXcoding/DocReal

4. Evaluation Benchmark

4.1. Limitation of DocUNet Benchmark

Despite being the most commonly used benchmark dataset, DocUNet [21] only contains 130 images with almost exclusively English text and limited document types, mainly consisting of receipts, magazines, and papers. Furthermore, the background of the distorted document images is relatively uniform and clean, and the types of deformations are not diverse enough, making them relatively ideal with clear document edges. Therefore, there is a need for a more diverse and comprehensive dataset to evaluate the performance of document image dewarping methods.

4.2. The Proposed DocReal Benchmark

To better evaluate the dewarping performance in real-life, we propose the first Chinese distorted image benchmark, called DocReal, which covers a wide range of real-life document image scenarios. DocReal includes 10 sub-scenarios across three significant scenarios: work, study, and daily life. The work scenario includes tables, contracts, and notices, while the study scenario includes papers, books, tests, and notes. The daily life scenario includes receipts, certificates, and newspapers. All these sub-scenarios cover the vast majority of usage scenarios in Chinese people's daily lives. To ensure the realism of the benchmark, we conducted desensitization photography, proportionate to the actual shooting environment and deformation type of the collected user data, with the user's authorization.

Each sub-scenario contains five different contents, and for each content, there are four distorted images with varying types of deformation, shooting angles, and shooting distances. The deformation types include curled, perspective, skew, and folded, commonly encountered in real-life document images. In total, DocReal contains 200 images within five classes, each representing a different deformation type (refer to Table 1) to facilitate qualitative comparisons, providing a comprehensive and diverse dataset for researchers in the field of document image dewarping. Detailed visualization is provided in Sec. C of the Supplementary.

Table 1. **Deformation types in the DocReal benchmark.**

Class	Deformation Type	Proportion
(a) Curled	Bending or curling	33%
(b) Skew	Slanted appearance	27%
(c) Perspective	Perspective warping	21%
(d) Folded	Visible creases	12%
(e) Flat	Inapparent distortion	7%

We believe that DocReal will be a valuable benchmark for the development and evaluation of document image rectification, providing a diverse range of real-life scenarios.

5. Experiments

We evaluate the effectiveness of our method in rectifying images through experiments on two benchmark datasets, DocUNet and the newly proposed DocReal. We provide qualitative and quantitative comparisons, assess the quality of rectification from a document analysis perspective, and conduct an ablation study to demonstrate the impact of Enet and attention modules on dewarping performance.

5.1. Experimental Setup

Evaluation Metrics. We use two evaluation criteria: image similarity and Optical Character Recognition (OCR) performance. For image similarity, we use Multi-Scale Structural Similarity (**MS-SSIM**) [32] and Local Distortion (**LD**) [36]. MS-SSIM is a multi-scale extension of SSIM as a weighted average of the SSIM index at different scales. LD measures the difference between the dewarped and reference images at each pixel location. For a fair comparison, all images are resized to a 598,400-pixel area and use the same weights of SSIM in [21]. For OCR performance, we use Edit Distance (**ED**) [18] and Character Error Rate (**CER**), providing complementary OCR information. ED measures the minimum number of insertions, deletions, and substitutions required to transform the OCR output into the ground truth, while CER measures the percentage of incorrectly recognized characters in the OCR output. All OCR metrics are evaluated on 60 images from DocUNet in accordance with [7–9] and all 200 images from DocReal using Tesseract v5.3.0, pytesseract v0.3.10, and "chi_sim_best" language file in an OCR engine on Windows 10 system.

5.2. Comparisons on the DocUNet Benchmark

Qualitative Comparison. Our method achieves high accuracy in restoring distorted text and images, resulting in clearer and more recognizable content, as shown in Fig. 5. Furthermore, our method is capable of handling a wide range of deformation types, which enables it to handle various real-world scenarios. Overall, the qualitative improvements provided by our method demonstrate its effectiveness in enhancing the visual quality of distorted documents and improving the accuracy of OCR.

Quantitative Comparison. Our method demonstrates superior performance compared to existing methods on the DocUNet benchmark [21], as shown in Table 2. Compared to DDCP [35], our method improves MS-SSIM and LD from 0.47 to 0.50 and 8.77 to 7.03, respectively, outperforming the majority of existing methods. We also achieve the lowest scores for ED and CER metrics among all methods, indicating the best OCR performance. These results demonstrate that our method is highly effective in addressing image distortion and enhancing OCR performance.

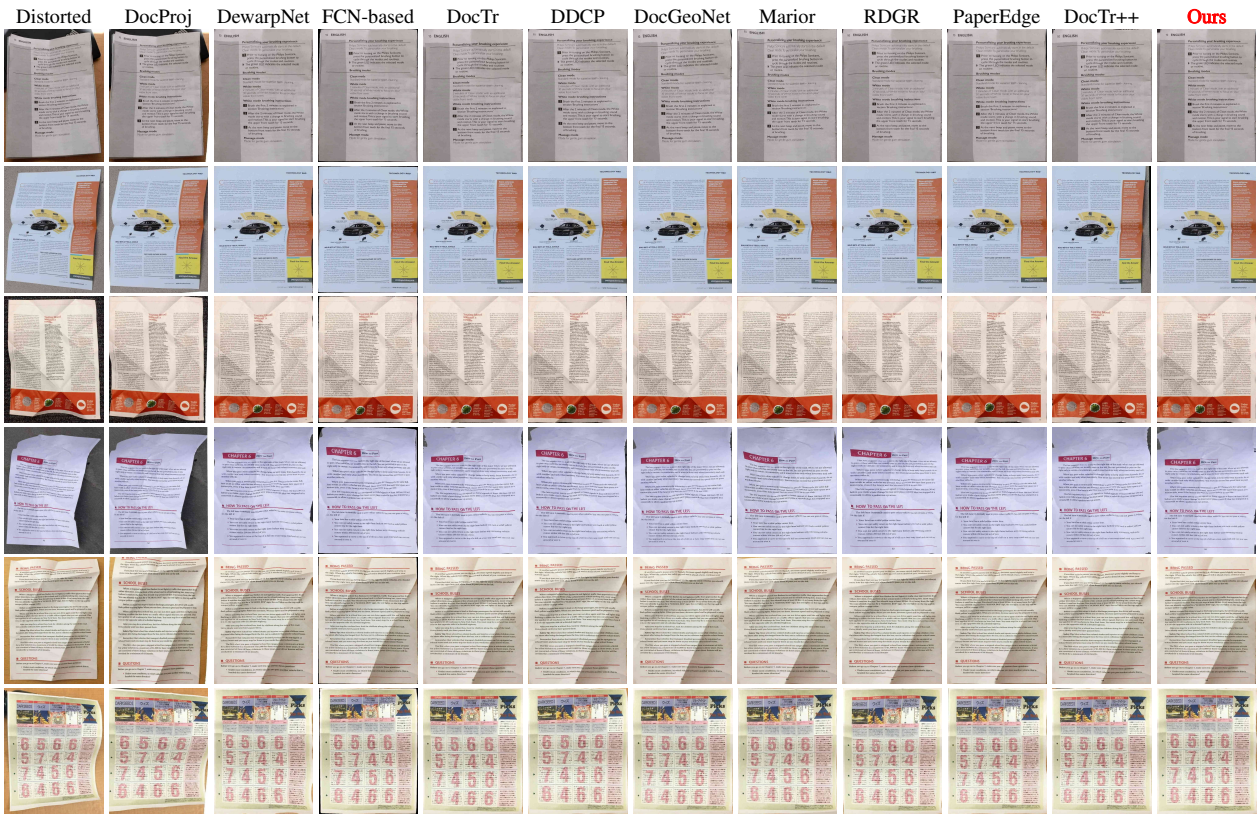


Figure 5. **Qualitative comparisons on the DocUNet [21] benchmark** with existing deep learning methods.

Table 2. **Quantitative comparisons on the DocUNet [21] benchmark.** "↑" indicates the higher the better and "↓" means the opposite. **Bold** font indicates best, and underline indicates second-best.

Method	MS-SSIM ↑	LD ↓	ED ↓	CER ↓
Distorted	0.25	20.51	1551.57	0.5087
DocUNet [21]	0.41	14.19	1259.83	0.3966
DocProj [19]	0.29	18.22	1353.30	0.4172
DewarpNet [5]	0.47	8.39	524.93	0.2100
FCN-based [34]	0.43	7.75	892.03	0.2916
DocTr [8]	**0.51**	7.76	464.17	0.1744
DDCP [35]	0.47	8.77	504.15	0.1811
DocGeoNet [9]	0.50	7.71	378.37	0.1506
Marior [37]	0.48	7.44	592.95	0.2132
RDGR [13]	0.50	8.51	419.63	0.1557
PaperEdge [20]	0.47	7.99	375.02	0.1538
DocTr++ [7]	0.46	9.37	508.27	0.1814
Ours	0.50	**7.03**	**359.90**	**0.1441**

5.3. Comparisons on the DocReal Benchmark

To evaluate the performance on real-life Chinese document images, we conduct tests on the DocReal benchmark. As shown in Fig. 6, our method demonstrates robust performance in handling various deformation types, including curled, skew, perspective, folded, and flat. This versatility makes our method well-suited for practical applications in real life. Additionally, our method produces text with improved readability and straightness, resulting in state-of-the-art performance across all metrics, as shown in Table 3.

Our results visually demonstrate a more complete text body than previous methods, despite the limitations of the MS-SSIM metric, which may not fully reflect the quality of dewarped images and is sensitive to line alignment [20]. Nevertheless, our approach outperforms existing methods in terms of visualization with improved readability.

Table 3. **Quantitative comparisons on the DocReal benchmark.** "↑" indicates the higher the better and "↓" means the opposite.

Method	MS-SSIM ↑	LD ↓	ED ↓	CER ↓
Distorted	0.32	35.79	698.37	0.6279
DocProj [19]	0.31	33.70	723.29	0.5840
DocTr [8]	0.55	12.66	455.12	0.4856
DDCP [35]	0.46	16.04	478.79	0.4688
DocGeoNet [9]	0.55	12.22	455.00	0.5051
PaperEdge [20]	0.52	11.46	420.78	0.4656
DocTr++ [7]	0.45	19.88	483.59	0.4978
Ours	**0.56**	**9.83**	**414.91**	**0.4582**

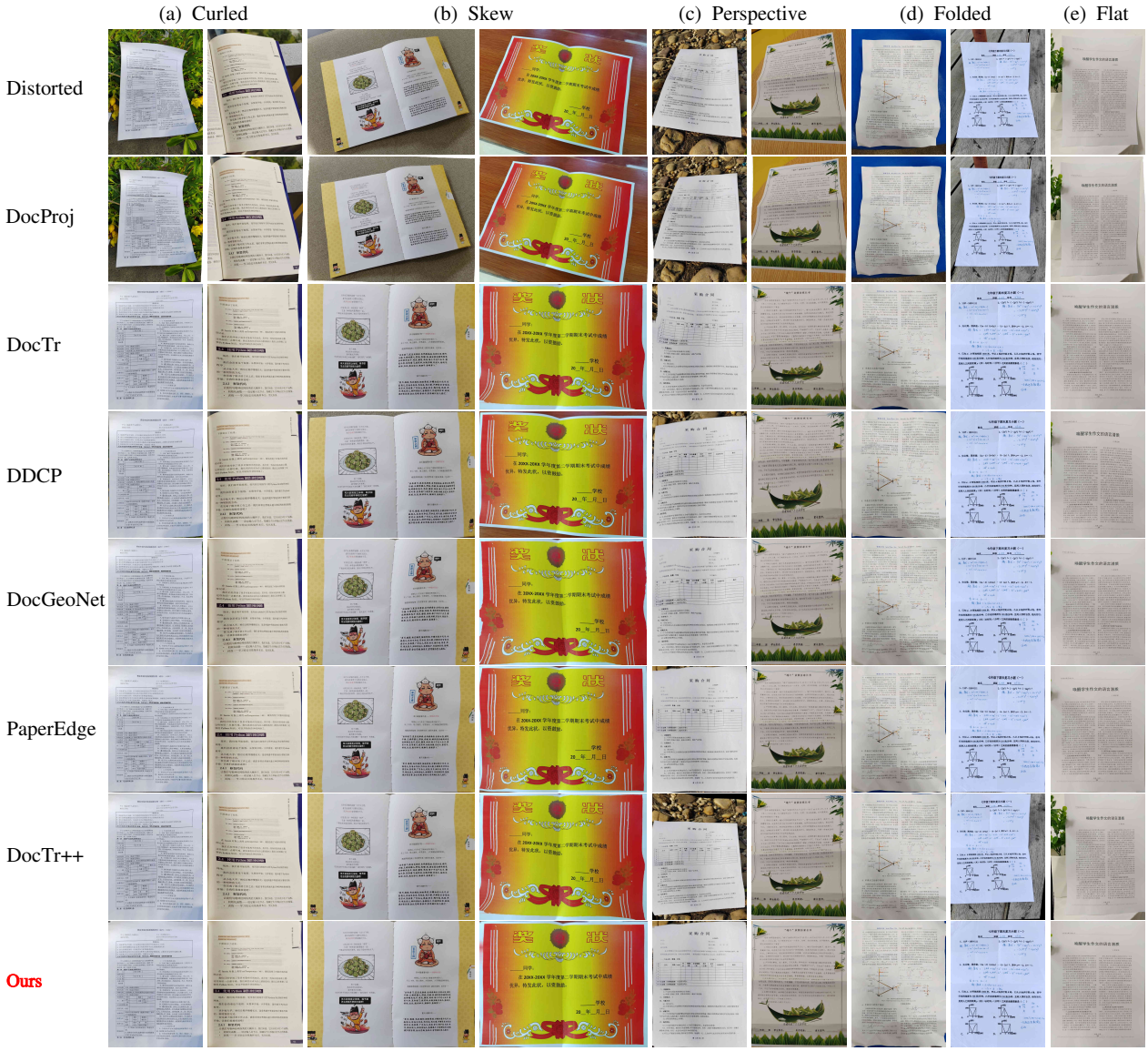
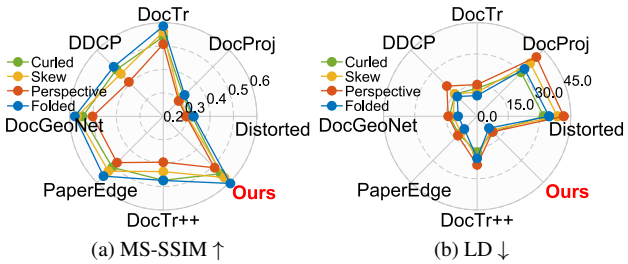


Figure 6. **Qualitative comparisons on the DocReal benchmark** with different deformation types using existing deep learning methods.

We further evaluate the robustness and compare the image similarities for four different deformations, as illustrated in Table 1. As shown in Fig. 7, our method consistently outperforms other methods in terms of the MS-SSIM and LD metrics. Compared to the DDCP [35], our method achieves comprehensive improvements of 13%, 25%, 25%, and 21% for the curled, skew, perspective, and folded deformation types, respectively, in terms of the MS-SSIM metric, and similar improvements of 29%, 39%, 49%, and 40% for four deformation types in terms of the LD metric. Numerical results are provided in Sec. D of the Supplementary.



(a) MS-SSIM ↑ (b) LD ↓

Figure 7. **Robustness comparisons on the DocReal benchmark** with four different deformation types."↑" indicates the higher the better and "↓" means the opposite. Our method achieves the best.

5.4. Case Study

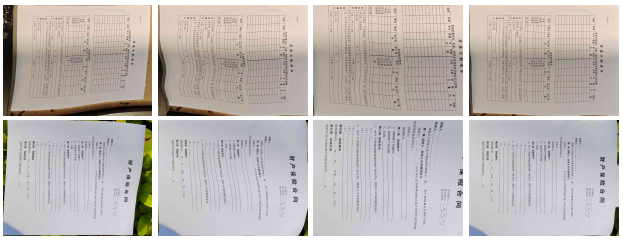
A case study, as shown in Fig. 8, reveals that while DocTr, DocGeoNet, and DocTr++ exhibit comparable results with our method in terms of the MS-SSIM metric, they still struggle to capture the document body when encountering disturbances, resulting in skewed text and reduced readability. More comparisons are in Sec. E of Supplementary.

DocTr DocGeoNet DocTr++ **Ours**

Figure 8. **Case study of background removal and readability.**

5.5. Ablation Study

We address the limitations of the DDCP method's residual backgrounds and Enet's potential to cause secondary deformation and reduced readability. As shown in Fig. 9, naïvely combining Enet with DDCP may result in worse readability due to inaccurate control point prediction. Thus, we combine Enet and attention-enhanced control point (AECP) module to arrive at a robust solution with well-captured document body and improved readability.



DDCP Enet Enet + DDCP **Ours**

Figure 9. **Our improvements on limitations of DDCP and Enet.**

Our ablation study focuses on two key factors: training data and model module. As shown in Table 4, we first demonstrate that using training data with 3D deformation (setting (b)) leads to significantly better performance compared to using 2D deformation (setting (a)). This is because 3D deformations enable us to generate large amounts of training data that closely resemble real-world deformation conditions, effectively enhancing the model's capabilities.

In terms of the model module, we compare setting (c) and setting (d) to show the improved performance brought by the Enet. On the other hand, the attention module enables the model to accurately capture local information, resulting in more effective local rectification with better OCR performance, when comparing setting (b) and setting (d). Detailed analysis is given in Sec. F of the Supplementary.

Overall, the ablation study highlights the importance of using 3D deformation types and incorporating Enet and attention modules to achieve the best results.

Table 4. **Ablation study on DocUNet and DocReal benchmarks.**

Module	Setting	(a)	(b)	(c)	(d)
Data	+ 2D type	✓			
	+ 3D type		✓	✓	✓
Model	+ Enet	✓	✓		✓
	+ attention			✓	✓
DocUNet	MS-SSIM ↑	0.31	0.49	0.36	**0.50**
	LD ↓	20.09	7.17	12.36	**7.03**
	ED ↓	593.63	394.37	639.65	**359.90**
	CER ↓	0.2387	0.1508	0.2194	**0.1441**
DocReal	MS-SSIM ↑	0.44	0.55	0.35	**0.56**
	LD ↓	21.18	9.99	29.33	**9.83**
	ED ↓	501.38	422.75	470.94	**414.91**
	CER ↓	0.5889	0.4775	0.4751	**0.4582**

6. Conclusion

We propose DocReal, a robust document dewarping approach for real-life images that effectively removes background noise using Enet and captures local deformations more accurately using an attention-enhanced control point (AECP) module. To improve the model's capabilities, we synthesize 2D images with 3D deformations and additional deformation types to augment the training data. We also provide a comprehensive benchmark of 200 Chinese distorted images with diverse real-life scenarios. Our approach achieves state-of-the-art performance on both the DocUNet and the newly proposed DocReal benchmark, and demonstrates superior dewarping accuracy, OCR performance, and robustness to various types of image distortion, making it a promising solution for real-life image dewarping.

Acknowledgments

The work of Fangchen Yu and Wenye Li was supported in part by Guangdong Basic and Applied Basic Research Foundation (2021A1515011825), Guangdong Introducing Innovative and Entrepreneurial Teams Fund (2017ZT07X152), Shenzhen Science and Technology Program (CUHKSZWDZC0004), and Shenzhen Research Institute of Big Data Scholarship Program. The work of Jianfeng Mao was supported in part by National Natural Science Foundation of China under grant U1733102, in part by the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen under grant B10120210117, and in part by CUHK-Shenzhen under grant PF.01.000404.

References

[1] Thierry Blu, Philippe Thévenaz, and Michael Unser. Linear interpolation revitalized. *IEEE Transactions on Image Processing*, 13(5):710–719, 2004. 3

[2] Michael S Brown and W Brent Seales. Document restoration using 3d shape. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9–12. Citeseer, 2001. 1

[3] Michael S Brown and W Brent Seales. Image restoration of arbitrarily warped documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(10):1295–1306, 2004. 1

[4] Michael S Brown and Y-C Tsoi. Geometric and shading correction for images of printed materials using boundary. *IEEE Transactions on Image Processing*, 15(6):1544–1554, 2006. 1, 4

[5] Sagnik Das, Ke Ma, Zhixin Shu, Dimitris Samaras, and Roy Shilkrot. Dewarpnet: Single-image document unwarping with stacked 3d and 2d regression networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 131–140, 2019. 1, 2, 4, 6

[6] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4268–4277, 2021. 1

[7] Hao Feng, Shaokai Liu, Jiajun Deng, Wengang Zhou, and Houqiang Li. Deep unrestricted document image rectification. *arXiv preprint arXiv:2304.08796*, 2023. 1, 2, 5, 6

[8] Hao Feng, Yuechen Wang, Wengang Zhou, Jiajun Deng, and Houqiang Li. Doctr: Document image transformer for geometric unwarping and illumination correction. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 273–281, 2021. 1, 2, 5, 6

[9] Hao Feng, Wengang Zhou, Jiajun Deng, Yuechen Wang, and Houqiang Li. Geometric representation learning for document image rectification. In *Proceedings of the European Conference on Computer Vision*, pages 475–492. Springer, 2022. 1, 5, 6

[10] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2

[12] Yuan He, Pan Pan, Shufu Xie, Jun Sun, and Satoshi Naoi. A book dewarping system by boundary-based 3d surface reconstruction. In *12th International Conference on Document Analysis and Recognition*, pages 403–407. IEEE, 2013. 1

[13] Xiangwei Jiang, Rujiao Long, Nan Xue, Zhibo Yang, Cong Yao, and Gui-Song Xia. Revisiting document image dewarping by grid regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4543–4552, 2022. 2, 6

[14] Taeho Kil, Wonkyo Seo, Hyung Il Koo, and Nam Ik Cho. Robust document image dewarping method using text-lines and line segments. In *14th IAPR International Conference on Document Analysis and Recognition*, volume 1, pages 865–870. IEEE, 2017. 1

[15] Beom Su Kim, Hyung Il Koo, and Nam Ik Cho. Document dewarping via text-line based optimization. *Pattern Recognition*, 48(11):3600–3614, 2015. 1

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[17] 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(7):1551–1562, 2009. 1

[18] Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics Doklady*, volume 10, pages 707–710. Soviet Union, 1966. 5

[19] 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(6):1–11, 2019. 2, 6

[20] Ke Ma, Sagnik Das, Zhixin Shu, and Dimitris Samaras. Learning from documents in the wild to improve document unwarping. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 1, 2, 3, 6

[21] Ke Ma, Zhixin Shu, Xue Bai, Jue Wang, and Dimitris Samaras. Docunet: Document image unwarping via a stacked u-net. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4709, 2018. 1, 2, 5, 6

[22] Amir Markovitz, Inbal Lavi, Or Perel, Shai Mazor, and Roee Litman. Can you read me now? content aware rectification using angle supervision. In *Proceedings of the European Conference on Computer Vision*, pages 208–223. Springer, 2020. 1

[23] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda Shapiro, and Hannaneh Hajishirzi. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 552–568, 2018. 3

[24] Erik Meijering. A chronology of interpolation: from ancient astronomy to modern signal and image processing. *Proceedings of the IEEE*, 90(3):319–342, 2002. 2

[25] Gaofeng Meng, Ying Wang, Shenquan Qu, Shiming Xiang, and Chunhong Pan. Active flattening of curved document images via two structured beams. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3890–3897, 2014. 1

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015. 3

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of Medical Image Computing and Computer Assisted Intervention*, pages 234–241. Springer, 2015. 2

[28] Yau-Chat Tsoi and Michael S Brown. Geometric and shading correction for images of printed materials: a unified ap-

proach using boundary. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–I. IEEE, 2004. 1

[29] Yau-Chat Tsoi and Michael S Brown. Multi-view document rectification using boundary. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 1

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[31] Floor Verhoeven, Tanguy Magne, and Olga Sorkine-Hornung. Neural document unwarping using coupled grids. *arXiv preprint arXiv:2302.02887*, 2023. 2

[32] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. IEEE, 2003. 5

[33] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, pages 3–19, 2018. 3

[34] Guo-Wang Xie, Fei Yin, Xu-Yao Zhang, and Cheng-Lin Liu. Dewarping document image by displacement flow estimation with fully convolutional network. In *Document Analysis Systems: 14th IAPR International Workshop, DAS 2020, Wuhan, China, July 26–29, 2020, Proceedings 14*, pages 131–144. Springer, 2020. 1, 2, 6

[35] Guo-Wang Xie, Fei Yin, Xu-Yao Zhang, and Cheng-Lin Liu. Document dewarping with control points. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16*, pages 466–480. Springer, 2021. 1, 2, 3, 4, 5, 6, 7

[36] Shaodi You, Yasuyuki Matsushita, Sudipta Sinha, Yusuke Bou, and Katsushi Ikeuchi. Multiview rectification of folded documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(2):505–511, 2017. 1, 5

[37] Jiaxin Zhang, Canjie Luo, Lianwen Jin, Fengjun Guo, and Kai Ding. Marior: Margin removal and iterative content rectification for document dewarping in the wild. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2805–2815, 2022. 2, 6

[38] Li Zhang, Yu Zhang, and Chew Tan. An improved physically-based method for geometric restoration of distorted document images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):728–734, 2008. 1