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

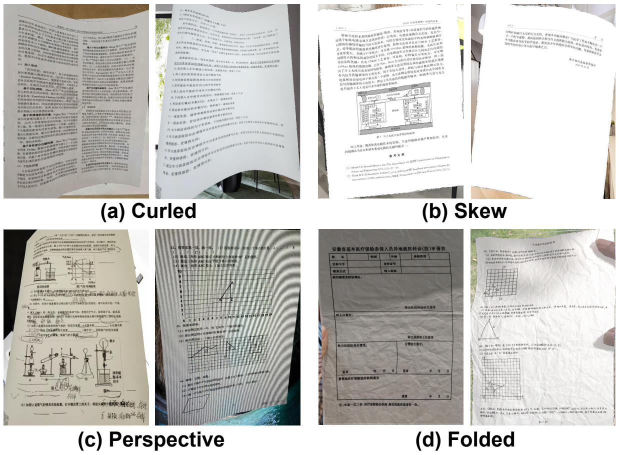
Fangchen Yu[1]*, Yina Xie[2], Lei Wu[2], Yafei Wen[2], Guozhi Wang[2]
Shuai Ren[2], Xiaoxin Chen[2], Jianfeng Mao[1,3], Wenye Li[1,3]†
[1]The Chinese University of Hong Kong, Shenzhen, China
[2]vivo AI Lab, Shenzhen, China
[3]Shenzhen Research Institute of Big Data, Shenzhen, China

This document serves as supplementary material, including

- **Sec. A** visualizes the synthetic training data;
- **Sec. B** provides the details of evaluations;
- **Sec. C** visualizes the proposed DocReal benchmark;
- **Sec. D** presents numerical results of robustness;
- **Sec. E** conducts a case study of the rectification;
- **Sec. F** conducts an ablation study of our method.

A. Synthetic Training Data

We synthesized 2D training data with 3D deformations, along with additional deformation types generated by our formulas. Furthermore, we added various types of noise to 2D images to augment the training data, such as moire patterns, fingerprints, shadows, and others, as shown in Fig. 1.



(a) Curled

(b) Skew

(c) Perspective

(d) Folded

Figure 1. **Synthetic training data with various types of noise.**

B. Evaluation Details

We evaluated the MS-SSIM and LD metrics on MATLAB R2023a using the same evaluation codes and SSIM weights in [2, 6]. We evaluated ED and CER metrics in an OCR engine on a Windows 10 system with Tesseract v5.3.0, pytesseract v0.3.10, and "chi_sim_best" language file[1].

C. Visualization of DocReal Benchmark

Fig. 2 visualizes the first Chinese distorted image benchmark, DocReal, including 200 images with 5 deformations.

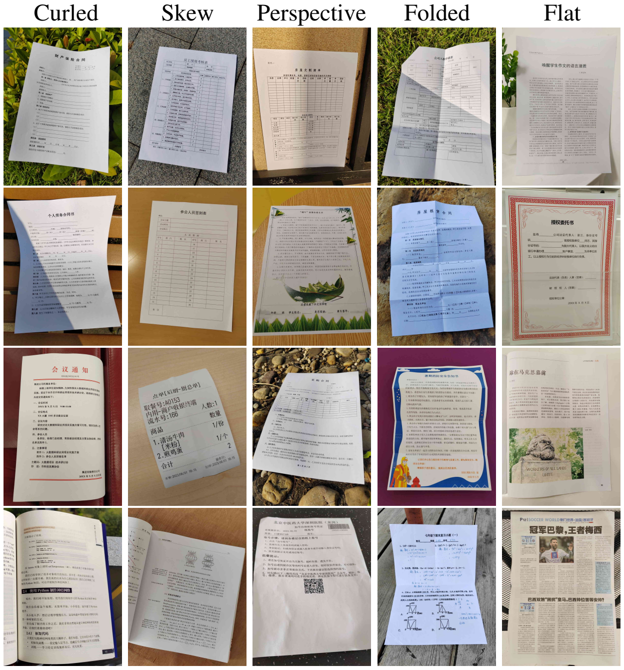


Figure 2. **Distorted images in the DocReal benchmark.**

*Work done during Fangchen Yu's internship at vivo AI Lab

†Corresponding author: wyli@cuhk.edu.cn

[1] https://github.com/tesseract-ocr/tessdata_best/blob/main/chi_sim.traineddata

D. Numerical Results

Table 1 summarizes the numerical results of the robustness comparisons, as reported in Fig. 7 of the main text. The results demonstrate the superior and robust performance of the proposed method on different types of deformations.

Table 1. **Quantitative comparisons of robustness on the DocReal benchmark**. "↑" indicates the higher the better and "↓" means the opposite. **Bold** font indicates the best.

Deformation	Curled	Skew	Perspective	Folded
Metric		MS-SSIM ↑		
Distorted	0.323	0.302	0.298	0.330
DocProj	0.322	0.299	0.293	0.329
DocTr	0.540	0.561	0.509	0.585
DDCP	0.483	0.456	0.408	0.500
DocGeoNet	0.545	0.569	0.502	0.578
PaperEdge	0.508	0.531	0.480	0.561
DocTr++	0.473	0.436	0.396	0.474
Ours	**0.547**	**0.570**	**0.511**	**0.605**
Metric		LD ↓		
Distorted	32.02	38.61	41.65	34.60
DocProj	29.88	36.10	40.28	32.20
DocTr	13.45	11.55	15.23	9.93
DDCP	15.07	15.28	20.56	13.46
DocGeoNet	13.32	11.49	13.89	9.14
PaperEdge	12.31	10.99	13.06	8.67
DocTr++	17.23	20.13	23.38	20.23
Ours	**10.62**	**9.36**	**10.55**	**8.06**

E. Case Study

As a reminder, our proposed method aims to address two significant issues: background residue and reduced readability. We conduct a detailed case study to better illustrate the limitations of existing image dewarping methods.

Limitations of Previous SOTA and Our Improvements. Despite having comparable MS-SSIMs with our method, previous state-of-the-art (SOTA) methods (i.e., DocTr [2], DocGeoNet [3], and DocTr++ [1]) still face challenges with residual background and reduced readability. Environmental noise, such as obstructions, similar background colors, and interference lines, may result in residual backgrounds, as illustrated in Fig. 3. Furthermore, these methods exhibit limited robustness for different text types, such as receipts (row 2 in Fig. 3), certificates (row 3 in Fig. 3), and tickets (row 5 in Fig. 3). As a result, they may produce dewarped images with severe deformations, further compromising their readability. Additionally, they also exhibit limited performance in terms of readability, as evidenced by curved tables (row 1 in Fig. 4), slanted text (rows 2-3 in Fig. 4), and twisted paper (rows 4-5 in Fig. 4). In comparison, our proposed method exhibits significant superiority in terms of background removal and text readability.

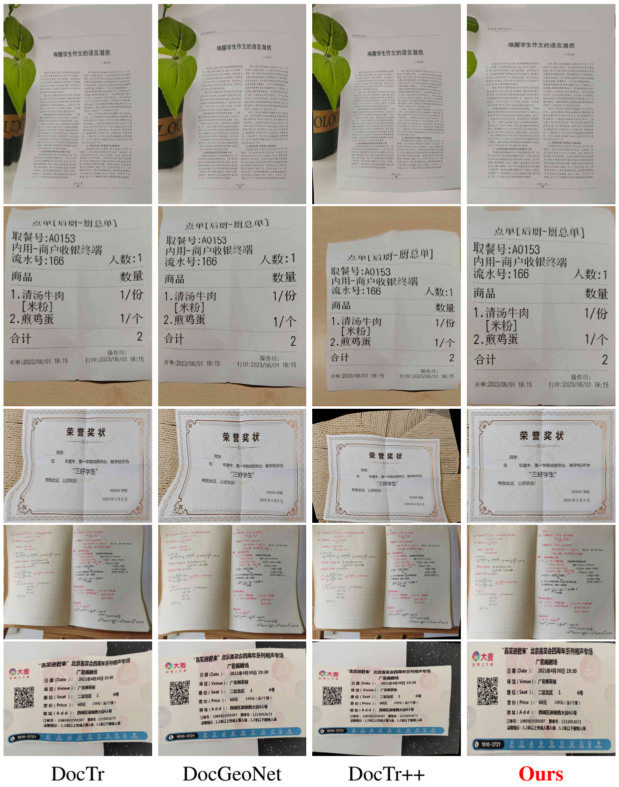


Figure 3. **Case study of background removal.**

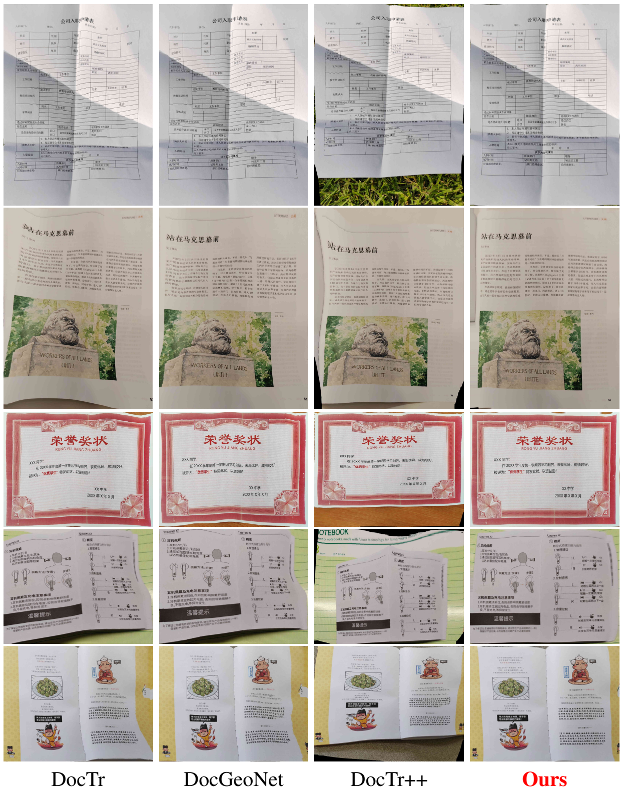


Figure 4. **Case study of text readability.**

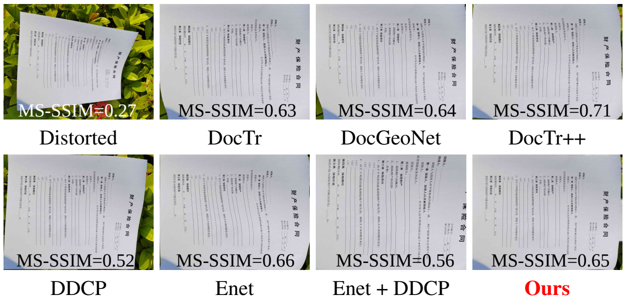
Differences Between Our Method and Existing Works. We address the limitations of the DDCP [9] method's residual backgrounds and Enet [5] network's reduced readability.

- Regarding the background removal, as shown in Fig. 6: The DDCP method aims to remove backgrounds by accurately predicting control points, capturing the document's body and deformation. However, in practice, due to various shooting environments, complex backgrounds, and diverse document types, predicting these control points is notably challenging, undermining its effectiveness and leaving some residual backgrounds. Meanwhile, simply merging Enet with DDCP might lead to suboptimal outcomes, producing overcropped images and omitting vital details due to imprecise control point prediction. In contrast, our approach predicts control points using preliminary results from the Enet method. This effectively addresses the issue and enhances prediction accuracy, as illustrated in Table 2.
- Regarding the text readability, as shown in Fig. 7: While Enet is proficient in background removal, it can inadvertently deform content, diminishing text readability. This might severely distort text and table lines. To mitigate this, we integrate Enet with the attention-enhanced control point (AECP) module, ensuring a more reliable solution that preserves the document's structure and enhances its readability.

Table 2. **Improvements of our method on the existing works.** "↑" indicates the higher the better and "↓" means the opposite.

Dataset	Method	MS-SSIM ↑	LD ↓	ED ↓	CER ↓
DocUNet	DDCP	0.47	8.77	504.15	0.1811
	Enet	0.47	8.08	760.52	0.2692
	Enet+DDCP	0.31	20.09	593.63	0.2387
	Ours	**0.50**	**7.03**	**359.90**	**0.1441**
DocReal	DDCP	0.46	16.04	478.79	0.4688
	Enet	0.52	11.32	526.24	0.5527
	Enet+DDCP	0.44	21.18	501.38	0.5889
	Ours	**0.56**	**9.83**	**414.91**	**0.4582**

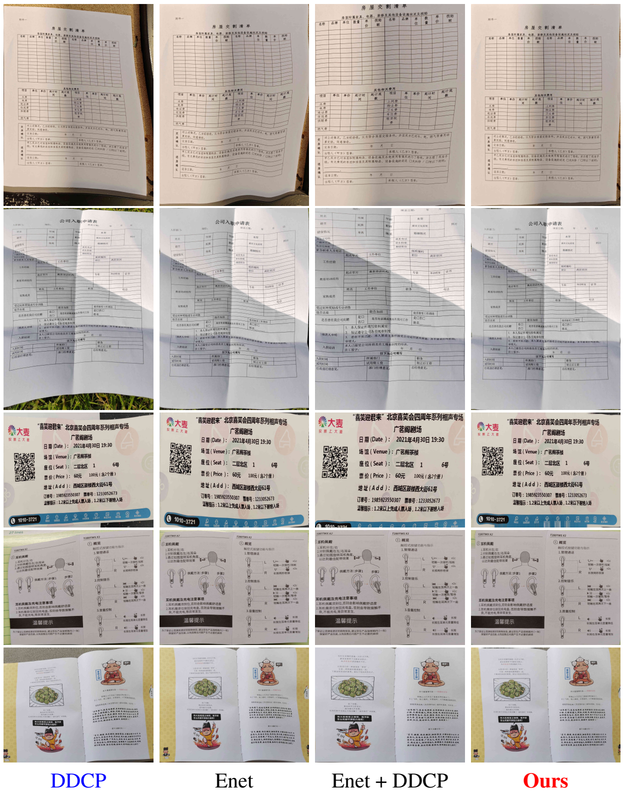
Limitation of the Evaluation Metric. Although our method produces comparable MS-SSIM results, it outperforms existing methods in terms of removing backgrounds and producing more readable images, as shown in Fig. 5.



Distorted DocTr DocGeoNet DocTr++

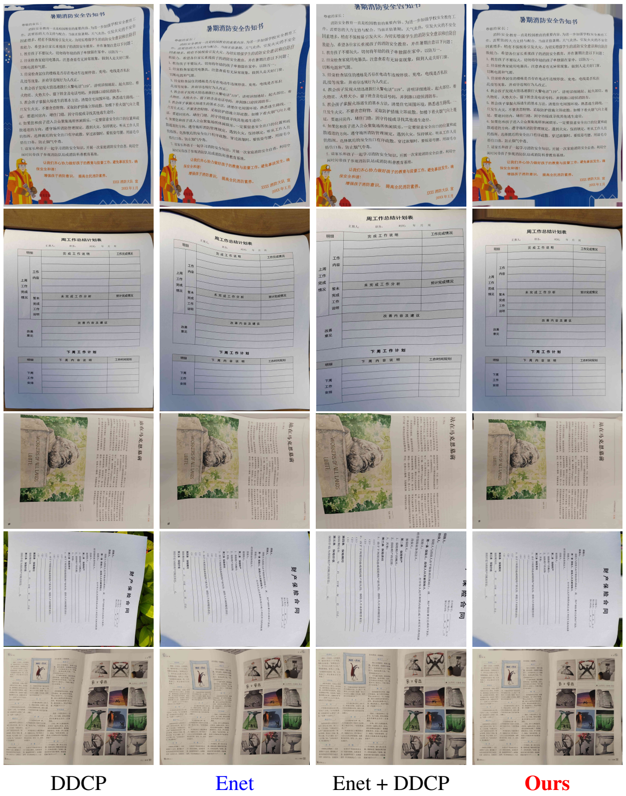
DDCP Enet Enet + DDCP Ours

Figure 5. **Limitation of MS-SSIM and improved visualization.**



DDCP Enet Enet + DDCP Ours

Figure 6. **Our improvements on background removal.**



DDCP Enet Enet + DDCP Ours

Figure 7. **Our improvements on text readability.**

F. Ablation Study

We conduct a detailed ablation study on the DocReal benchmark to evaluate the effectiveness of our proposed method. Our model, which incorporates the Enet architecture and attention-enhanced control point (AECP) module, features four submodules, as discussed in the main text. Due to limited computation resources, we focus on the contributions of three modules: Enet, attention, and dilation pyramid. Ablation results on the DocUNet and DocReal benchmarks are presented in Table 3.

Our study brings to light the following insights:

(a) Excluding Enet leads to subpar outcomes due to residual backgrounds. Moreover, without Enet, the efficacy of the attention module is hampered, yielding inaccurate document structures.

(b) Using the preliminary output from Enet, the attention module markedly boosts OCR performance, reflecting in improved ED and CER metrics. This is achieved by both text flattening and precise document structure recognition.

(c) The dilation pyramid module is pivotal in enhancing our model's effectiveness, impacting image similarities and OCR performance. This potent technique allows the network to harness multi-scale features, making it highly adept for imaging tasks [4, 8].

By weaving together Enet with the attention-enhanced control point (AECP) module, our model delivers superior dewarping outcomes in both image similarities and OCR performance, as illustrated in Figs. 8 and 9, respectively.

Scientific Justification of Model Designs: Our design choices, grounded in both empirical results and prior research, yield an architecture optimized for performance, which we elucidate below.

1) Strategic Positioning of Attention Modules: We strategically position the attention module post-shallow features (green in Fig. 2 of the main text) to accentuate texture and lighting nuances. An additional attention module post-deep features (blue in Fig. 2 of the main text) serves to underline text lines and global deformities. We plan to conduct further ablation experiments of locations of attention modules in the future.

2) Channel and Spatial Attention: For insights into the channel attention's positioning, [8] offers clarity. Moreover, the Convolutional Block Attention Module (CBAM) paper [7] elucidates the "C & S" sequence relationships, further validating our choices.

3) Efficacy of the Dilation Pyramid: We have rigorously tested the merits of dilated convolution with a dilation pyramid, as shown in Table 3. Our findings unambiguously indicate the pronounced advantages of this sub-module, reinforcing its inclusion in our design.

4) Sub-module Foundations: For a thorough understanding of the foundational aspects of the green, blue, and orange sub-modules elucidated in the main text, [8] serves as a comprehensive reference.

Table 3. **Ablation study of modules on the DocUNet and DocReal benchmarks.** "↑" indicates the higher the better and "↓" means the opposite. **Bold** font indicates best.

Dataset		DocUNet		DocReal	
Module	Setting	MS-SSIM ↑	LD ↓	MS-SSIM ↑	LD ↓
Enet	**w/** (d)	**0.50**	**7.03**	**0.56**	**9.83**
	w/o (a)	0.36	12.36	0.35	29.33
Attention	**w/** (d)	**0.50**	**7.03**	**0.56**	**9.83**
	w/o (b)	0.49	7.17	0.55	9.99
Dilation	**w/** (d)	**0.50**	**7.03**	**0.56**	**9.83**
	w/o (c)	0.47	7.90	0.54	10.30

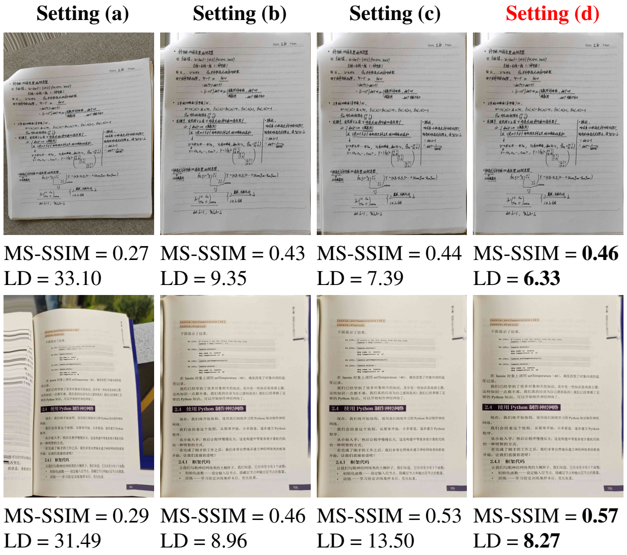


Figure 8. **Ablation study on image similarities.**

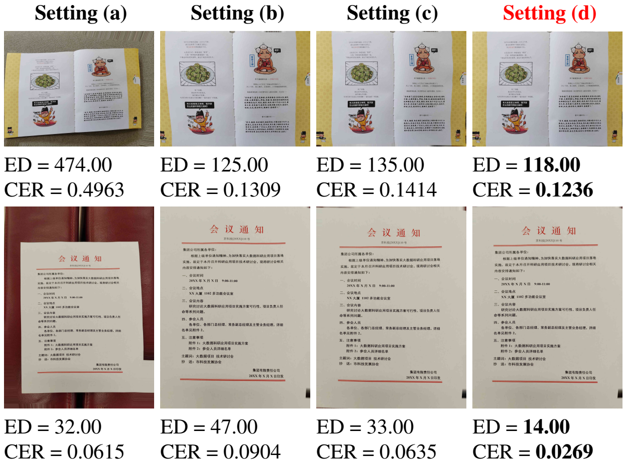


Figure 9. **Ablation study on OCR performance and readability.**

References

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

[2] 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

[3] 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. 2

[4] Mourad Gridach. Pydinet: Pyramid dilated network for medical image segmentation. *Neural Networks*, 140:274–281, 2021. 4

[5] 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. 3

[6] 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

[7] 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. 4

[8] 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. 4

[9] 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. 3