Geometric Representation Learning for Document Image Rectification (Supplementary Material)

Hao Feng[1], Wengang Zhou[1,2*], Jiajun Deng[1],
Yuechen Wang[1], and Houqiang Li[1,2*]

[1] CAS Key Laboratory of GIPAS, EEIS Department,
University of Science and Technology of China
{fh1995,wyc9725}@mail.ustc.edu.cn, {zhwg,dengjj,lihq}@ustc.edu.cn
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

1 Comparison with Prevalent Software

Technically, the prevalent document rectification algorithms in smartphones commonly have a restriction that the document should be a regular quadrilateral shape. Specifically, such techniques first detect the four corner points of the document to localize a quadrilateral document region and then apply perspective transformation to get the rectified image. As a result, they can not deal with the situation when the captured document has any irregular deformations.

As shown in Figure 1, we compare our method with the prevalent software, including the CamScanner Application, the internal document rectification algorithm of IPhone 12, Huawei Nova 9, and Xiaomi 11. We can see that our DocGeoNet is capable of rectifying the documents with irregular deformations. This is because the predicted warping flow of DocGeoNet defines a non-parametric transformation, thus being able to represent a wide range of distortions.

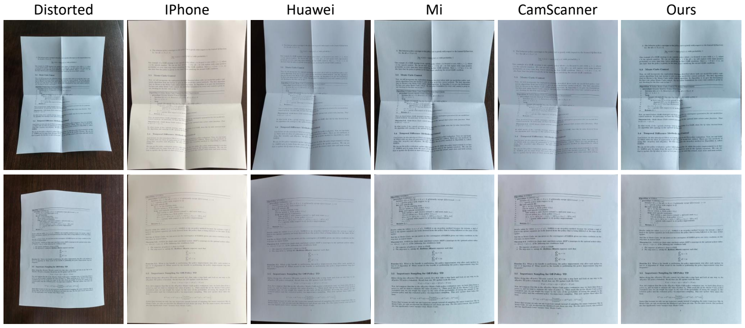


Fig. 1. Qualitative comparisons of our method with the prevalent software, including the CamScanner Application, the internal document rectification algorithm in smartphone of IPhone, Huawei, and Xiaomi.

[*] Corresponding Authors: Wengang Zhou and Houqiang Li.

2 More Qualitative Results

As shown in Figure 2, we present more qualitative rectified results on the DocUNet Benchmark dataset [25]. Besides, as shown in Figure 5, we provide more rectified results on real distorted document photos. As we can see, the proposed DocGeoNet shows superior rectification quality.

Particularly, as shown in Figure 5, the distorted images show various physical deformations, backgrounds, and illumination conditions. These photos are captured under various indoor (during day and night) and outdoor scenes. Besides, the used documents contain text, tables, figures, or their mixture.

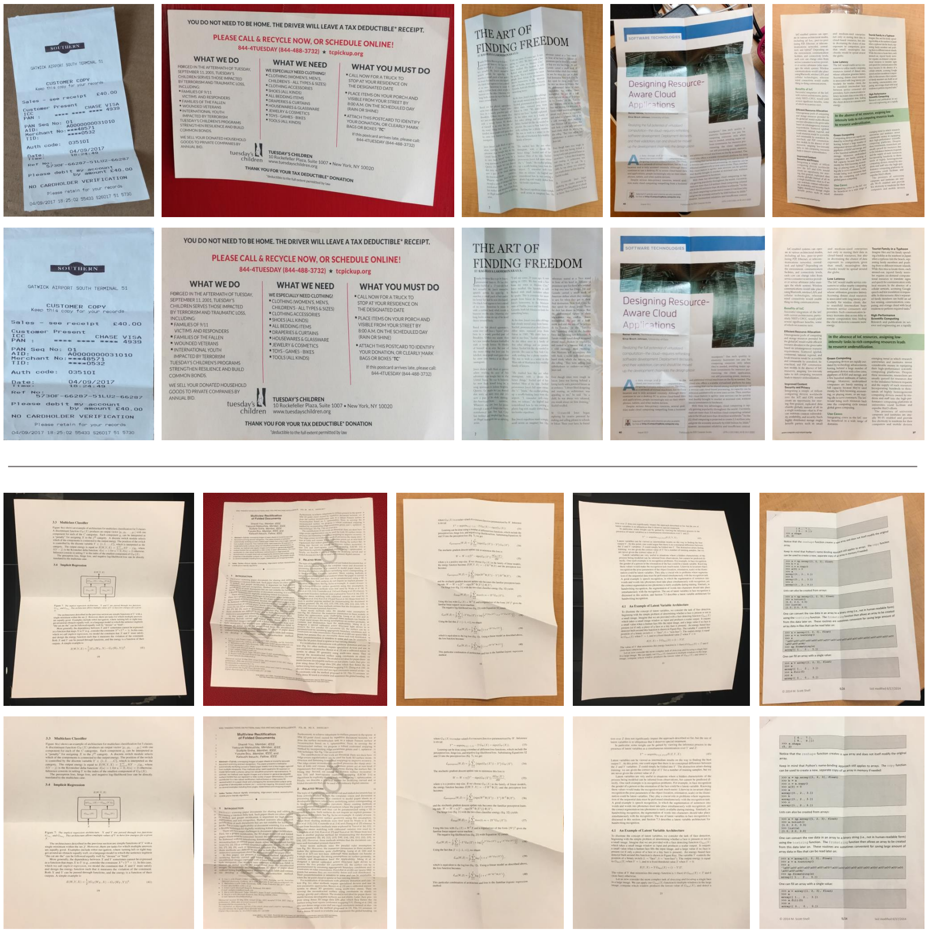


Fig. 2. Visualization of the rectified results on the DocUNet Benchmark dataset [25]. The first and third row are the input distorted images. The second and bottom row show their corresponding rectified results.

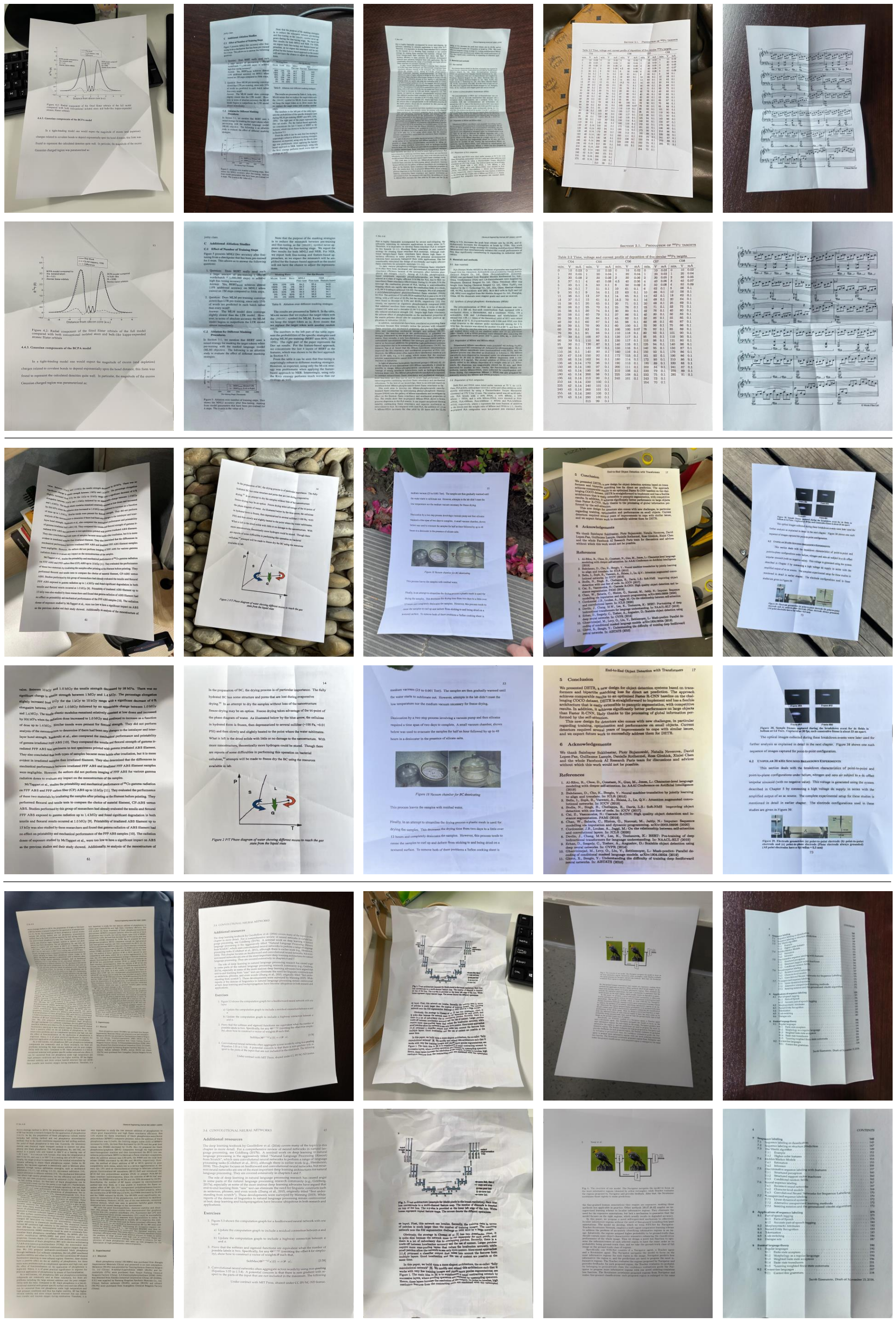


Fig. 3. Visualization of the rectified results on real distorted document photos under various conditions. The first, third and fifth row are the distorted images, and the remaining rows are their rectified results.

3 DIR300 Dataset

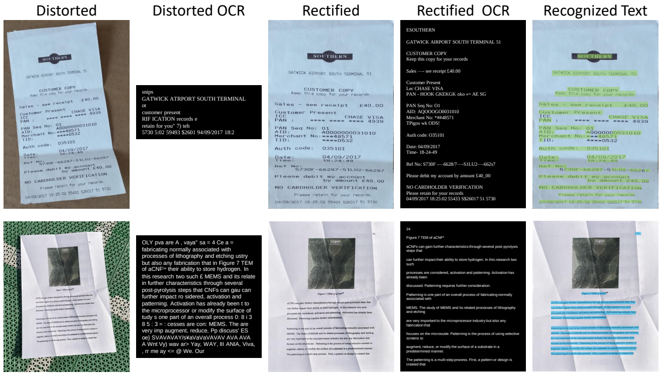
Furthermore, we present the detail about the creation of the DIR300 test set. Concretely, the images are firstly taken by three people with three different cellphones, including iPhone 12, Huawei Nova 9, and Mi 11. Each person captures 100 photos. Secondly, to involve various backgrounds and illumination conditions, for each person, the indoor and outdoor scenes account for 70% and 30%, respectively. For indoor scenes, 20% are taken in the evening. Thirdly, in terms of distortions, 40% samples involve random curving; 40% samples contain random folds; 10% samples are flat; the remaining 10% are heavily crumpled. Here we do not fix the scenes for a certain distortion.

Note that the ground truth images are captured before the collection of the distorted images. Specifically, we put the regular rectangular document on a plane. Then, we align the four corner points of the rectangular document and then get a perfect rectification. The rectified image is taken as the ground truth. Another way is adopted by the successful DocUNet Benchmark dataset [25] which scanned the printed document to image as GT, but a perfect alignment is still difficult due to the scanning error.

As shown in Figure 5, we present more qualitative rectified results on the DIR300 test set. It can be seen that, the proposed DocGeoNet shows superior rectification performance.

4 OCR Visualizations

To reveal the impact of the geometric rectification on the OCR performance, we further visualize their OCR results, respectively. We use the Tesseract [33] as the OCR engine. As shown in Figure 4, after the geometric rectification, the OCR performance makes remarkable improvements.



```
snips
GATWICK ATRPORT SOUTH TERMINAL
ot
customer present
RIF ICATION records e
retain for you" 7) teh
5730 5:02 59493 $2601 94/09/2017 18:2
```

```
ESOUTHERN

GATWICK AIRPORT SOUTH TERMINAL 51

CUSTOMER COPY
Keep this copy for your records

Sales —- see receipt £40.00

Customer Present
Lec CHASE VISA
PAN - HOOK GKEKGK oko «= AE SG

PAN Seq No: O1
AID: AQOOOGO0031010
Merchant No: *##48571
TPigos wk ODS!

Auth code: O35101

Date: 04/09/2017
Time- 18-24-49

Ref No: S730F -—6628/7—-S1LU2—-662s7

Please debit my account by amount £40_00

NO CARDHOLDER VERIFICATION
Please retain for your records
04/09/2017 18:25:02 55433 S$26017 51 5730
```

```
OLY pva are A , vaya" sa = 4 Ce a =
fabricating normally associated with
processes of lithography and etching ustry
but also any fabrication that in Figure 7 TEM
of aCNF™ their ability to store hydrogen. In
this research two such £ MEMS and its relate
in further characteristics through several
post-pyrolysis steps that CNFs can gau can
further impact ro sidered, activation and
patterning. Activation has already been t to
the microprocessor or modify the surface of
tudy s one part of an overall process 0: 8 i 3
8 5 : 3 = : cesses are con: MEMS. The are
very imp augment, reduce, Pp discuss' ES
oe} SVAVAVAY/s¥aVaVaVAVAV AVA AVA
A Wnt Vy) wav ar> Yay, WAY, Ill ANIA, Viva,
, rr me ay <= @ We. Our
```

```
24

Figure 7 TEM of aCNF*

aCNFs can gain further characteristics through several post-pyrolysis
steps that

can further impact their ability to store hydrogen. In this research two
such

processes are considered, activation and patterning. Activation has
already been

discussed. Patterning requires further consideration.

Patterning is one part of an overall process of fabricating normally
associated with

MEMS. The study of MEMS and its related processes of lithography
and etching

are very important to the microprocessor industry but also any
fabrication that

focuses on the microscale. Patterning is the process of using selective
screens to

augment, reduce, or modify the surface of a substrate in a
predetermined manner.

The patterning is a multi-step process. First, a pattern or design is
created that
```

Fig. 4. Visualization of two instances about the impact of the geometric rectification on the OCR performance. The second and fourth column show the recognized text of the distorted image and the rectified image of the proposed DocGeoNet, respectively. Besides, we highlight the correct recognized text of DocGeoNet in the fifth column.

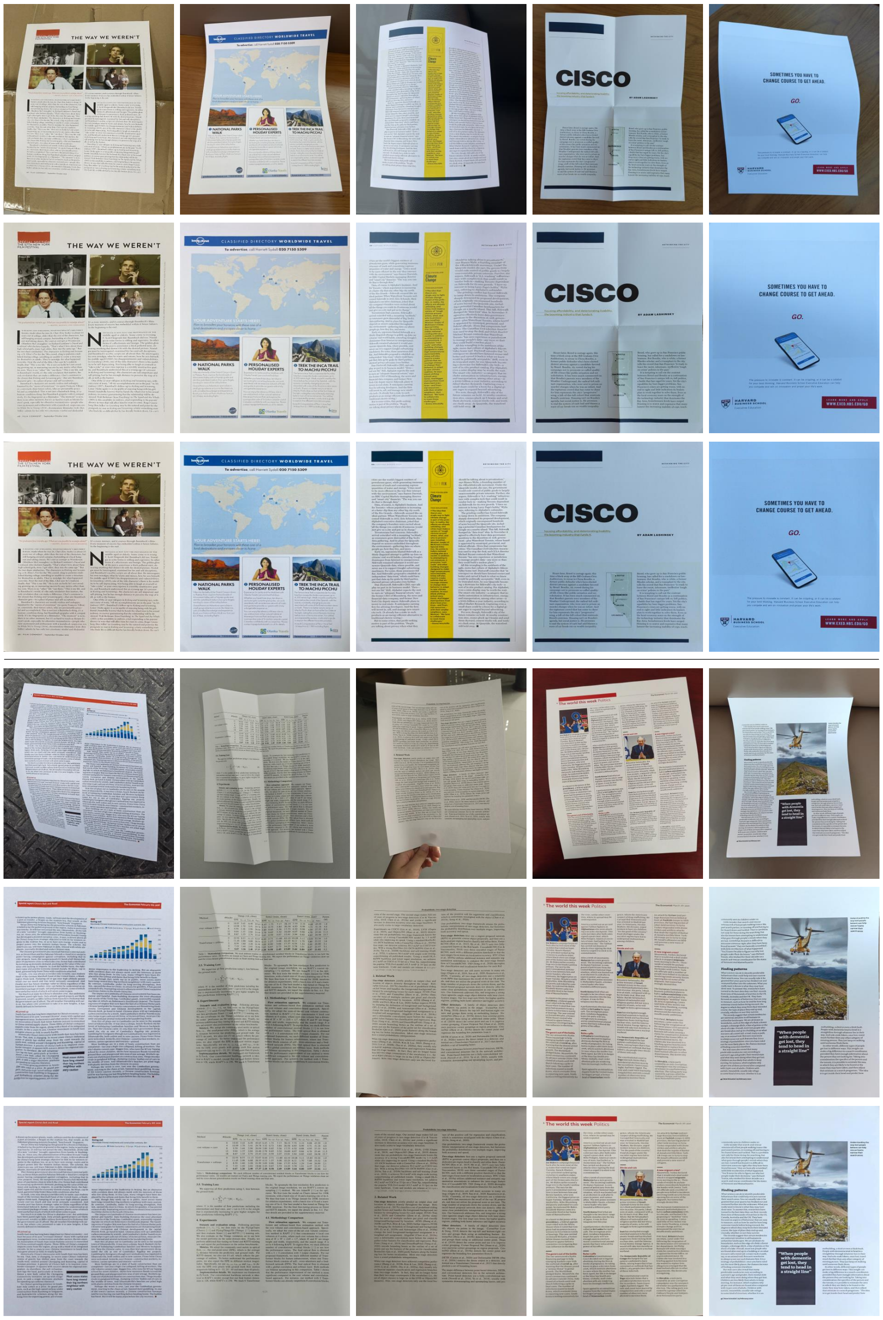


Fig. 5. Visualization of the rectified results on the DIR300 test set. The first and fourth row are the distorted images. The second and fifth row are their rectified results. The remaining are their corresponding ground truth.

5 Efficiency Analysis

In this section, we discuss the efficiency of the proposed DocGeoNet and other learning-based methods. As shown in Table 1 in the manuscript, DocGeoNet is efficient in terms of inference time and parameter count compared with existing learning-based methods. One of reasons is that we predict the backward warping flow directly that is used to sample the pixels from the input distorted image for rectification, following [6,9,10]. In contrast, previous work DocUNet [25] predicts the forward warping flow instead, which has to be converted to the backward warping flow first using the unstructured points. In addition, DocProj [19] crops the distorted image into patches first and then rectifies the patches to perform rectification. However, the rectification of input distorted patches and the stitching of backward warping flow patches heavily increase the computational cost.

6 Performance on DocUNet Benchmark

In Table 1 of the manuscript, we report the performance on the corrected DocUNet Benchmark dataset [25]. In this section, for clarity, we also report the results on the DocUNet Benchmark dataset [25] with two mistaken image samples. The results are shown in Table 1. Note that the two mistaken images are not contained in the sub-set for the OCR evaluation. Hence, they only affect the evaluation of MS-SSIM and LD.

Table 1. Quantitative comparisons of the existing learning-based methods in terms of image similarity, distortion metrics, OCR performance, and running efficiency on the DocUNet Benchmark dataset [25] **with two mistaken image samples**. "↑" indicates the higher the better, while "↓" means the opposite.

Methods	Venue	MS-SSIM ↑	LD ↓	ED ↓	CER ↓	FPS ↑	Para.
Distorted	-	0.2464	20.51	2111.56/1552.22	0.5352/0.5089	-	-
DocUNet [25]	*CVPR'18*	0.4094	14.22	1933.66/1259.83	0.4632/0.3966	0.21	58.6M
AGUN [22]	*PR'18*	0.4491	12.06	-	-	-	-
DocProj [19]	*TOG'19*	0.2928	18.19	1712.48/1165.93	0.4267/0.3818	0.11	47.8M
FCN-based [42]	*DAS'20*	0.4361	8.50	1792.60/1031.40	0.4213/0.3156	1.49	23.6M
DewarpNet [6]	*ICCV'19*	0.4692	8.98	885.90/525.45	0.2373/0.2102	7.14	86.9M
PWUNet [7]	*ICCV'21*	0.4879	9.23	1069.28/743.32	0.2677/0.2623	-	-
DocTr [9]	*MM'21*	0.5085	8.38	724.84/464.83	0.1832/0.1746	7.40	26.9M
DDCP [41]	*ICDAR'21*	0.4706	9.51	1442.84/745.35	0.3633/0.2626	**12.38**	**13.3M**
FDRNet [43]	*CVPR'22*	**0.5440**	8.75	794.54/514.90	0.2010/0.1846	-	-
RDGR [14]	*CVPR'22*	0.4929	9.11	693.38/420.25	**0.1654**/0.1559	-	-
Ours	-	0.5027	**8.37**	**692.86**/**379.00**	0.1797/**0.1509**	-	24.8M

7 Metric Analysis

During the experiments, we find that the SSIM [38] is not a very appropriate metric for document image rectification. Document image rectification is not a pixel-aligned task, different them the typical pixel-aligned tasks, such as deraining and denoising. SSIM [38] and MS-SSIM [39] are designed to capture the perceptual distortion of images with respect to a reference image. Blur, noise, color shifts, and halos are the types of artifacts they are designed to capture, rather than geometric distortion (i.e., misalignment of the pixels between a reference and a corrupted image). For future works, we recommend removing it and leaving the other metrics. If the authors think they should keep it to be consistent with previous work, we recommend adding a note that that is the purpose of providing that number. A typical example is FDRNet [43]. The SSIM score in Table 1 in the manuscript is smaller than that in Table 1 in this supplementary material. However, Table 1 in the manuscript shows the performance on the corrected DocUNet Benchmark dataset [25].

8 Limitation Discussion

Existing methods all limit the size of background area of the distorted images when training the rectification networks. This is because the background area of distorted images in DocUNet Benchmark dataset [25] is small. As a result, when increasing the camera distance, the background area becomes larger and the performance drops. It is the same with our method. We hope future works can propose more robust methods.