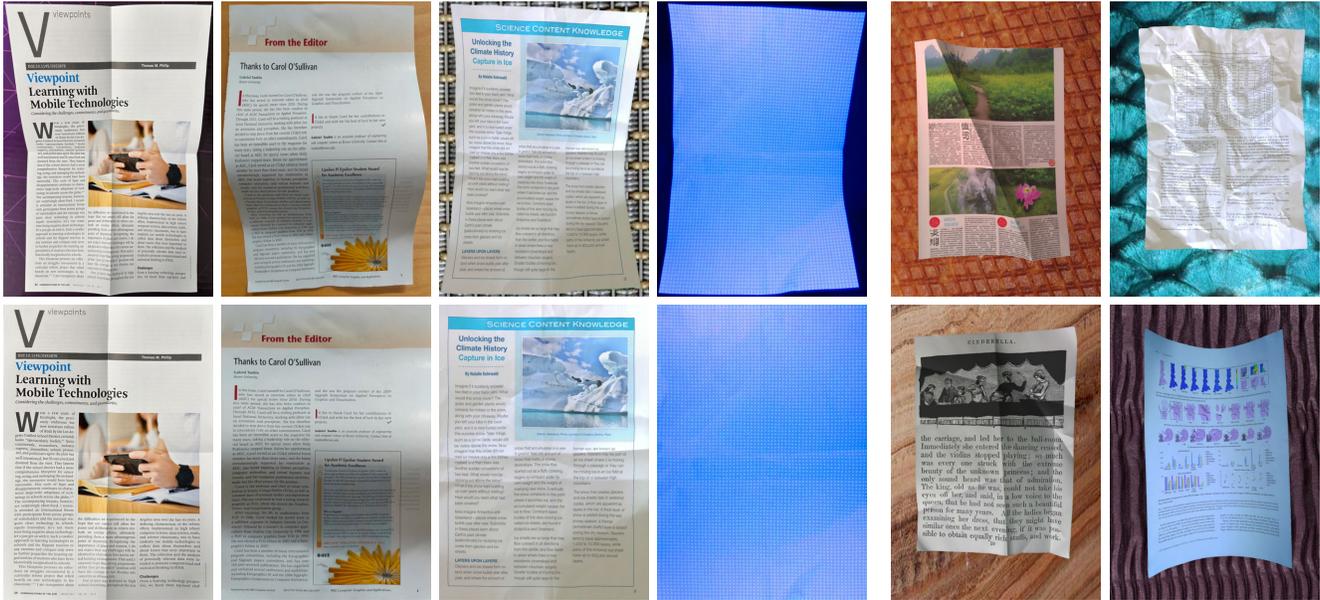


UVDoc: Neural Grid-based Document Unwarping

Floor Verhoeven
ETH Zurich
Switzerland
floor.verhoeven@inf.ethz.ch

Tanguy Magne
ETH Zurich
Switzerland
tanguy.magne@inf.ethz.ch

Olga Sorkine-Hornung
ETH Zurich
Switzerland
sorkine@inf.ethz.ch



Input images (top) and our unwarping (bottom)

Samples from our UVDoc dataset

Figure 1: Unwarping results and data samples from our UVDoc dataset. The first two columns are examples where the input images were captured on a mobile phone. The third column is a sample from our UVDoc dataset and the fourth shows the corresponding UV-lit image along with its unwarping (obtained based on the unwarping grid predicted for the image in the third column). The last two columns show a few more examples from our UVDoc dataset.

ABSTRACT

Restoring the original, flat appearance of a printed document from casual photographs of bent and wrinkled pages is a common everyday problem. In this paper we propose a novel method for grid-based single-image document unwarping. Our method performs geometric distortion correction via a fully convolutional deep neural network that learns to predict the 3D grid mesh of the document and the corresponding 2D unwarping grid in a dual-task fashion, implicitly encoding the coupling between the shape of a 3D piece of paper and its 2D image. In order to allow unwarping models to train on data that is more realistic in appearance than the commonly used synthetic Doc3D dataset, we create and publish our own dataset, called UVDoc, which combines pseudo-photorealistic document images with physically accurate 3D shape

and unwarping function annotations. Our dataset is labeled with all the information necessary to train our unwarping network, without having to engineer separate loss functions that can deal with the lack of ground-truth typically found in *document in the wild* datasets. We perform an in-depth evaluation that demonstrates that with the inclusion of our novel pseudo-photorealistic dataset, our relatively small network architecture achieves state-of-the-art results on the DocUNet benchmark. We show that the pseudo-photorealistic nature of our UVDoc dataset allows for new and better evaluation methods, such as lighting-corrected MS-SSIM. We provide a novel benchmark dataset that facilitates such evaluations, and propose a metric that quantifies line straightness after unwarping. Our code, results and UVDoc dataset are available at <https://github.com/tanguymagne/UVDoc>.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SA Conference Papers '23, December 12–15, 2023, Sydney, NSW, Australia

© 2023 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0315-7/23/12.

<https://doi.org/10.1145/3610548.3618174>

CCS CONCEPTS

- Computing methodologies → Computer vision; Image manipulation;
- Applied computing → Document capture.

KEYWORDS

Document unwarping, document dewarping, rectification, distortion correction, OCR, datasets

ACM Reference Format:

Floor Verhoeven, Tanguy Magne, and Olga Sorkine-Hornung. 2023. UVDoc: Neural Grid-based Document Unwarping. In *SIGGRAPH Asia 2023 Conference Papers (SA Conference Papers '23)*, December 12–15, 2023, Sydney, NSW, Australia. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3610548.3618174>

1 INTRODUCTION

The task of physical document digitization, e.g. for financial administration, is increasingly being done in a casual setting with the use of smartphones rather than the more traditional in-office flatbed scanners. However, the appearance of these casually captured images usually differs greatly from flatbed-scans due to varying camera angles, unconstrained illumination conditions and physical deformations of the paper, such as folding, crumpling and curving. These appearance variations pose a problem for post-processing steps, such as optical character recognition (OCR). Document image rectification is therefore an important step in the modernized document digitization pipeline, making layout extraction and OCR performance comparable to that of the traditional pipeline.

Several research efforts have been undertaken to tackle the problem of document unwarping using either model- or data-driven approaches. The model-driven approaches typically try to fit a constrained, piecewise-smooth surface to the imaged document. This geometric optimization is generally slow and has limited approximation capabilities, making it unsuitable for practical applications. Data-driven approaches instead rely on training a neural network to perform the unwarping. These methods are fast at runtime but typically require a large amount of high-quality training data, which can be difficult to obtain. The available training data can roughly be categorized as either synthetic or *in the wild* document images. The former group is generated by rendering images using 3D scans of real document geometries, whilst the latter simply consists of photographs of actual deformed documents. The synthetic category has the problem that dense 3D capture is often noisy, and rendering photorealistic paper can be challenging, making the appearance of the generated data samples non-realistic as a result. The challenge with the latter category is that ground truth data, most notably the ground truth unwarping function, is difficult to obtain.

Our main contribution is UVDoc, a dataset that aims to decrease the domain gap between the synthetic Doc3D dataset [Ma et al. 2018] usually used to train models for the task of document unwarping, and real document photographs. Our dataset contains 20,000 pseudo-photorealistic images of documents, and is equipped with all the required information to train a coarse grid-based document unwarping neural net. We achieve photorealistic appearance by using image compositing instead of rendering, thereby retaining the shading and material appearance from the original image capture. As our dataset is tailored to a coarse grid-based approach, it is easy to produce even though it includes numerous ground-truth annotations. We offer a new benchmark dataset whose rich ground-truth annotations allow for evaluation of the unwarping performance without the entanglement of shading artifacts, as well as a new

metric that measures the straightness of lines in the unwrapped image.

We train a small deep convolutional neural network that performs document image unwarping from a single RGB image. Its design is chosen specifically to make use of our UVDoc dataset. It uses a dual-head approach to predict both a 3D grid mesh representing the 3D shape of the document, as well as a 2D unwarping grid representing a coarse backward map. The backward mapping acts as an inverse parameterization; it indicates at each output pixel, which pixel coordinates should be sampled from the input image to produce the unwrapped image. This dual-task approach encodes an implicit coupling between the 2D and the 3D grid, just like there is a physical coupling between the 3D document shape and its 2D image. Since we learn a coarse 2D unwarping grid instead of a dense unwarping flow, our network size is greatly reduced compared to state-of-the-art methods.

Using our own relatively small model, and training on a combination of the large Doc3D synthetic dataset and our own custom UVDoc data, we obtain state-of-the-art performance on the DocUNet benchmark for most evaluation criteria. Moreover, we show that the addition of our UVDoc dataset improves the performance of existing document unwarping methods.

2 RELATED WORK

Document image unwarping is a widely studied topic. We divide previous work into two categories: model-based and data-driven approaches.

2.1 Model-based document unwarping

Early works take a geometric modeling approach and try to unwrap document images by first creating a 3D reconstruction of the document surface, which is then flattened onto the plane by solving an optimization problem. These works commonly obtain an estimate of the 3D document surface with the help of auxiliary equipment, such as structured light [Brown and Seales 2001, 2004], two structured laser beams [Meng et al. 2014] or laser range scanners [Zhang et al. 2008]. Other model-based methods use multi-view images instead of hardware to estimate the 3D shape of the document surface [Koo et al. 2009; Luo and Bo 2022; Tsoi and Brown 2007; Ulges et al. 2004; Yamashita et al. 2004; You et al. 2018]. Finally, Tian and Narasimhan [2011] exploit the structure of the document, such as lines, to reconstruct the 3D geometry through a shape-from-texture method.

Once the 3D reconstruction of the document surface is in place, different methods are used to flatten it to the plane. Brown and Seales [2001; 2004] and Zhang et al. [2008] flatten the document surface using a simulation of a stiff mass-spring system falling down to a plane under gravity. Another common technique is to fit a (piecewise-)smooth parametric surface to the estimated 3D document surface and flatten it according to a parameterization. This approach can involve generalized cylinders [Kil et al. 2017; Kim et al. 2015; Koo et al. 2009; Meng et al. 2018; Nachappa et al. 2023; Zhang et al. 2004], generalized ruled surfaces [Meng et al. 2014; Tsoi and Brown 2007], smooth developable surfaces [Liang et al. 2005, 2008] and NURBS [Yamashita et al. 2004; Zhang and Tan 2005].

Parametric approaches often heavily rely on the texture flow of the text lines in the document to estimate parametric line directions, making them less suitable for documents that only contain sparse text. Additionally, their optimization-based nature makes these methods slow and unsuitable for real-time applications, and their dependence on auxiliary equipment makes their use in real-world scenarios inconvenient and costly.

2.2 Data-driven document unwarping

Data-driven document unwarping methods work directly on a single RGB image of a document, employing deep learning to infer a 2D displacement fields or a coarse grid that can be used to unwarped the distorted input image. Ma et al. [2018] are one of the first to propose such a network, using two chained U-Nets to predict the forward mapping (the first estimates an initial guess, and the second refines it). DewarpNet [Das et al. 2019] also employs two chained networks inferring first the 3D coordinates and then the backward mapping from the 3D coordinates. They create the large synthetic Doc3D dataset with rich annotations to make this possible. Xu et al. [2022] build on this approach, but use siamese losses to additionally encourage deformation and texture consistency.

Several other ideas have been implemented. Patch-based methods [Das et al. 2021; Li et al. 2019] predict the displacement field independently on different parts of the image, thus better handling local distortions at the cost of having to properly stitch the different patches together. Iterative methods [Feng et al. 2021b; Zhang et al. 2022] progressively refine the predicted warping flow field and predict a foreground segmentation mask before starting the iterative rectification process, removing the burden of localizing the document boundaries from the unwarping network. Several methods based on textlines also use foreground segmentation. Jiang et al. [2022] use these pieces of information as explicit constraints of an optimization problem. Feng et al. [2022] feed a concatenation of textlines and 3D shape features into a network that predicts the displacement map. Recent work by Das et al. [2022] learns a texture parameterization for neural representations through differentiable rendering, using multi-view input in a data-driven approach.

A variety of different network architectures have been proposed, such as fully convolutional neural networks [Xie et al. 2020], pyramid encoder-decoder networks [Liu et al. 2020], or transformers [Feng et al. 2021a]. Other works use transformer architectures to tackle specific use cases, such as partially visible documents [Feng et al. 2023] or invoices [Hertlein et al. 2023].

Recently works by Xie et al. [2021], Xue et al. [2022] and Ma et al. [2022] follow the approach of predicting a coarse backward mapping. Some of these are capable of learning from images captured *in the wild*, either by direct comparison of the Fourier-filtered unwarped images [Xue et al. 2022], or by designing a specific loss on pairs of slightly perturbed images [Ma et al. 2022].

Our dual-task-based network architecture enforces the model to predict physically plausible shapes and unwarping grids. It processes input images in a single stage without any segmentation pre-processing and predicts a coarse backward mapping rather than a dense displacement field, making it very efficient.

2.2.1 Datasets. The datasets used for training the methods mentioned above can be split into two categories; real and synthetic,

Table 1: Comparison between the different document unwarping datasets. The last column indicates whether the ground-truth Backward Mapping (BM) between the distorted and the unwarped document is available.

Dataset	# Samples	Type	BM
Doc3D [Das et al. 2019]	100,000	Synthetic	✓
DIW [Ma et al. 2022]	5,000	Real	✗
WarpDoc [Xue et al. 2022]	1,020	Real	✗
Ours	20,000	Pseudo-real	✓

with the latter being most commonly used. Earlier works [Ma et al. 2018; Xie et al. 2020, 2021] use synthetic datasets generated based on non-physically plausible 2D deformations. More accurate are datasets based on 3D deformations [Das et al. 2019; Li et al. 2019], such as Doc3D. Most recent works are trained using Doc3D or its variations with richer annotations [Feng et al. 2022], or cropped images [Feng et al. 2023]. Even though the 3D shapes in Doc3D are more realistic than those in [Ma et al. 2018], since they are based on depth captures of actual deformed papers, they are heavily smoothed compared to the original document shapes. The rendered appearance is also not very photorealistic, which causes performance degradation when using the network on actual photographs. In contrast, our UVDoc dataset, which is made from real, captured sheets of paper, is more realistic both visually and geometrically.

Datasets of real photographs of deformed documents, paired with their flatbed scans ground truth, have recently gathered increased interest. Some of these datasets provide segmentation information [Ma et al. 2022] but most do not come with further annotations [Xue et al. 2022]. These datasets are closer in appearance to real document images, but since they are equipped with very few annotations, they require the design of custom loss functions to train models. In comparison, our dataset is equipped with a lot of annotations while being visually similar to *in-the-wild* data. See comparative summary in Table 1.

3 THE UVDOC DATASET

We create our own dataset, UVDoc, containing 20k pseudo-photorealistic images of warped documents. Our motivation is to obtain a dataset of photorealistic document images that has more ground truth information available than *document in the wild* images, and more realistic appearance than synthetically generated renderings. This allows for a stronger supervision signal than what is available for general *document in the wild* data and benefits from more realistic appearance. We compare the main characteristics of our dataset against other available datasets in Table 1.

Capture. We print regular grids of dots, with grid size of 89×61 , on A4-sized pieces of paper using an inkjet printer with UV ink that is invisible to an RGB camera in regular light, but becomes visible in UV light in an otherwise dark room. We opt for this grid aspect ratio to obtain an equally spaced grid in both horizontal and vertical direction, and to approximate the aspect ratio of A4 paper in portrait mode, the most common paper type that documents are printed on. Note that on the paper boundary, we deviate slightly from a perfectly regular grid by offsetting the border dots a little,

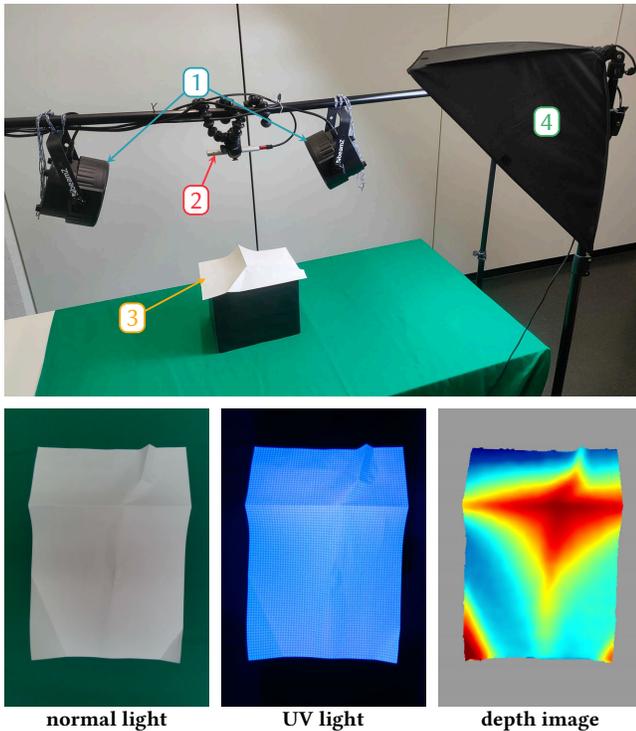


Figure 2: An overview of our data capture setup and sample data acquired in the process. The top shows our capture setup: [1] UV lights, [2] SR305 depth camera, [3] deformed sheet of paper, [4] regular light. The bottom shows a capture sample including RGB images of the normally lit and UV-lit paper, and its depth image.

so that they fully fit on the paper and can be detected more easily. We fold and bend the pieces of paper in various ways to emulate common deformations. We then capture pairs of RGB-D images of deformed papers using the Intel RealSense SR305 depth camera: one image in regular lighting and one in UV lighting (Fig. 2). We use two commercially available 30 W, 395 nm UV lamps and one 72 W, 395 nm UV lamp to reduce the amount of shadows in the UV-lit image. We also use a regular light with adjustable color temperature and brightness to create varying lighting conditions. We control the camera and the lights using a laptop and remote switches, so that there is no movement between the two captured frames, and the depth and pixel information is aligned. We capture various types of deformed paper, such as curved, folded, and crumpled, and we also vary the lighting conditions. The dataset contains a total of 1008 distinct geometries, which we augment to 4032 geometries by applying horizontal and/or vertical flips to each sample.

Recovering the grid. Using the UV-lit image, where the printed grid is visible, we obtain the pixel coordinates of the grid points on the deformed piece of paper. To detect them we use OpenCV’s SimpleBlobDetector, coupled with manual annotation for extreme cases where the automatic detection fails (less than 0.5% of the points need to be manually annotated). Once all points have been detected, we compute their correspondences to the vertices of

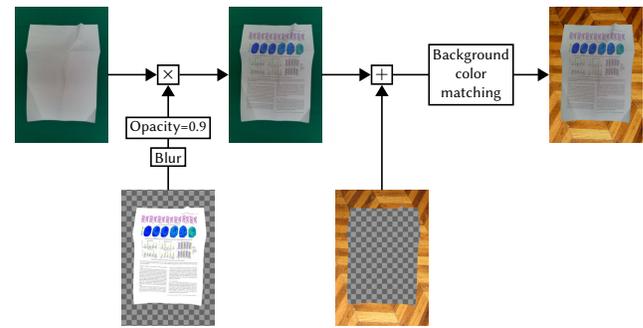


Figure 3: The pipeline used to create a sample of our UVDoc dataset. It combines the captured image of a blank paper, the texture and the background.

a regular grid, which is equivalent to ordering them as an 89×61 grid. The technical details of solving this ordering problem are described in the supplementary material.

We call the ordered grid the 2D unwarping grid. Combining the coordinates of the 2D unwarping grid with the depth values at these same pixel coordinates and the intrinsics of the camera, we construct a 3D grid mesh corresponding to the 3D shape of the piece of paper.

Pseudo-photorealistic image generation. Since we have a known mapping between the 2D unwarping grid and the original regular 2D grid, we can construct a coarse uv -parameterization of the 3D grid mesh. We use bilinear interpolation for the uv -parameterization when applying a texture to the geometry and for the 2D unwarping grid when performing the unwarping, to obtain a full-resolution dense backward mapping.

The uv -parameterization is used to apply a document texture on top of the image of the blank warped paper. The document textures include books and scientific articles sampled from the web, as well as other types of documents such as magazines, invoices, and music sheets, generated using a text-to-image model [AI 2023]. As illustrated in Fig. 3, we blend the document texture with the lighting-baked blank document image by multiplying the two images. This gives a pseudo-photorealistic combination between the lighting and the texture. We also replace the background in the image with a background sampled from the *Describable Textures* dataset [Cimpoi et al. 2014]. Finally, we apply color correction to match the hue of the background to the hue of the document and we also equalize the brightness of the background to the foreground. Using this approach, we create a dataset of 20,000 images in total. We provide the original lighting-baked blank document images along with the uv -parameterization, so users of the dataset can easily replace the document and the background textures if desired.

At the end of our data capture pipeline, we are equipped with a ground-truth 2D unwarping grid, a uv -parameterization and a 3D grid mesh for each sample in our dataset. Since we use the *physical* uv -parameterization recorded via the 2D grid, rather than a parameterization designed by a rendering engine, our texture gets deformed and applied with greater physical accuracy. Additionally, by circumventing a rendering pipeline, our images *look like real*

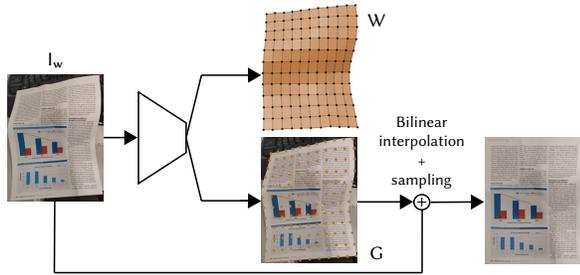


Figure 4: Our unwarping pipeline. We start with an RGB image of a warped document and feed it into our encoder-style network. The network predicts both a 3D grid mesh (top branch), as well as a 2D unwarping grid (bottom branch) in parallel. The 2D unwarping grid is then bilinearly interpolated to the desired output image resolution and is used to sample pixels from the input image to obtain the final unwarped document image.

paper, which is hard to simulate when rendering. The full UVDoc dataset is available at <https://github.com/tanguymagne/UVDoc>.

4 METHOD

To completely unwarped a document, we assume that the input photograph is taken from a camera position in which the document’s 3D shape can be represented as a height field, i.e., the entire document is visible and there are no occlusions and foldovers.

We use a dual-head network to predict a 45×31 2D unwarping grid G containing pixel coordinates, and a 45×31 grid mesh of 3D shape coordinates W from a warped 488-by-712 input image I_w . We do not predict G at the full ground-truth resolution in an attempt to keep our network as compact as possible. As illustrated in Fig. 4, the 2D unwarping grid G encodes the deformation that leads to the unwarped document: grid-point $G_{i,j}$ holds the pixel coordinates (relative to the image, in the range $[-1, 1]$) of the pixel that will be placed at position (i, j) in the unwarped image (up to constant scaling). The grid G can also be seen as a coarse backward mapping. Finally, G is bilinearly interpolated to the original image size. This upsampled backward mapping is used to generate the full-resolution unwarped image. The 3D grid mesh W is not used for the unwarping, but we incorporate it in training with an L_1 loss as a regularization term. This helps the network understand the underlying geometry of the document and improve the unwarping performance (see ablations studies in Sec. 6.2 and Table 5).

4.1 Network architecture

We use a relatively straightforward dual-head, fully convolutional encoder architecture inspired by the encoder part of the architecture used in [Xie et al. 2020]. The input image goes through two convolutional downsampling layers that each use a 5×5 kernel and reduce the image size by a factor of two. This is followed by three dilated residual blocks, which lead to a spatial pyramid with stacked dilated convolutions. Finally, two heads with two convolutional layers predict G and W , respectively. We give a detailed graphical overview of our architecture in the supplemental.

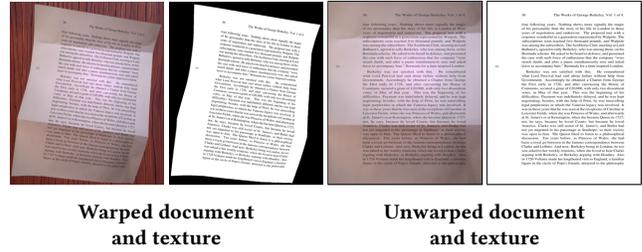


Figure 5: The shaded and unshaded version of a sample from the UVDoc benchmark, identical up to shading. The shaded version and unshaded version have a CER of 0.439 and 0.004 respectively and ED of 959 and 14. Note that the unshaded version has non-zero CER and ED as it is compared to the original texture, while it has been warped and unwarped using our coarse bilinearly interpolated 2D grid, and thus includes some artifacts.

4.2 Training loss

We denote the ground-truth variables as their regular symbols (e.g., G) and their predicted counterparts with a hat (e.g., \hat{G}). Our training loss is a combination of L_1 losses on both the 2D unwarping grid G and the 3D grid mesh W , as well as an image reconstruction loss \mathcal{L}_r :

$$\mathcal{L} = \alpha \|G - \hat{G}\|_1 + \beta \|W - \hat{W}\|_1 + \gamma \mathcal{L}_r, \quad (1)$$

where α, β, γ are weights used to balance the influence of the individual loss terms.

\mathcal{L}_r is an L_1 loss between the ground truth unwarped image and the image unwarped using the predicted unwarping grid G . For Doc3D samples, the reconstruction loss is computed directly on the unwarping of the input image I_w , which includes shading. For our UVDoc samples, to allow the network to focus on the content of the document rather than on shading artifacts, we compute the reconstruction loss using the unwarping of the unshaded document and the ground truth document texture. We provide further training details for our method in the supplementary material.

5 EVALUATION METRICS AND THE UVDoc BENCHMARK

We discuss common existing evaluation metrics for document unwarping and propose a new benchmark (UVDoc), along with a new metric that provides faithful evaluation even in the presence of varied shading.

5.1 UVDoc benchmark

To foster more detailed evaluation of document unwarping methods in the future, we create the UVDoc benchmark dataset. This benchmark is generated in a similar fashion to the UVDoc dataset but contains other geometries, document textures and backgrounds, not included in the main dataset. The benchmark consists of 50 images. Thanks to our pseudo-photorealistic data generation pipeline, we have access to pairs of warped images with and without lighting (and thus shadows) baked in, see Fig. 5. This setup provides new opportunities for meaningful metrics. For each sample in the benchmark, an unwarping pipeline can predict the unwarping function

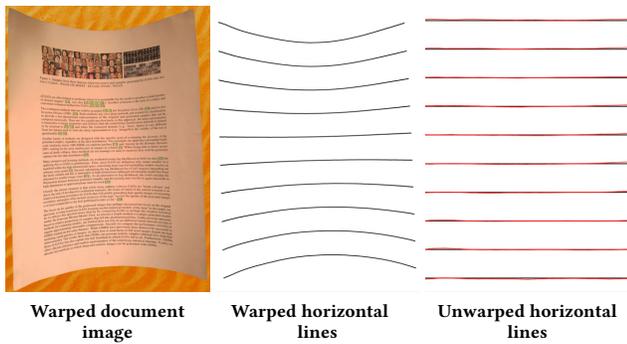


Figure 6: Our new horizontal line metric is the standard deviation of the y coordinate of warped horizontal lines (middle) unwarped using the predicted backward mapping (right in red, ground-truth in black).

for the shaded image, and then apply it to the unshaded image. The unwarped unshaded image can then be compared to the ground-truth original image texture. This way the effect of illumination can be removed and the unwarping deformation can be better evaluated.

5.2 Evaluation metrics

In our objective evaluations, we employ image similarity metrics as well as optical character recognition (OCR) performance. Following [Ma et al. 2018] and [Das et al. 2019], we use multi-scale structural similarity (MS-SSIM) and local distortion (LD) as metrics for the image similarity evaluation. We also employ the aligned distortion (AD) metric (introduced by [Ma et al. 2022]), which corrects some of the flaws of the previous metrics. We evaluate OCR performance using the character error rate (CER) and edit distance (ED).

As OCR engines are typically targeted towards use with images originating from flatbed scanners, they are ill-suited for text recognition on images with lighting variations and shadows [tesseract-ocr 2023]. The two images in Fig. 5 are identical up to shading but result in vastly different OCR performances. We therefore want to point out that OCR performance on the DocUNet dataset should be interpreted with care, since its baked-in shading plays such a large role. More details regarding our evaluation metrics can be found in the supplementary material.

5.2.1 Line straightness metric. Our UVDoc benchmark is annotated with not just a ground-truth unwarped image but also the ground-truth warping function, which allows us to design a new metric that evaluates the straightness of lines in the unwarped image. We generate triplets of images consisting of a warped document image and two images containing warped horizontal and vertical lines. These three images are generated using the same geometry and thus correspond to the same ground-truth unwarping function. We can now predict the unwarping function from the warped document image, and then apply it to the warped line images, giving us the unwarped horizontal and vertical lines. A perfectly predicted unwarping function should map the lines (which are 1 pixel thick) back to exactly horizontal and vertical lines. By measuring the average standard deviation of the lines we obtain a measure of how well the unwarping function maps horizontal (resp. vertical) lines

Table 2: Quantitative unwarping performance comparisons on the DocUNet benchmark dataset. Bold font indicates best, underline indicates second-best and *italic* indicates third-best score. The last column compares the network sizes, expressed in number of parameters (millions). We compare our results against DewarpNet [Das et al. 2019], DispFlow [Xie et al. 2020], DocTr [Feng et al. 2021a], PW Unwarping [Das et al. 2021], DDCP [Xie et al. 2021], FDRNet [Xue et al. 2022], RDGR [Jiang et al. 2022], Marior [Zhang et al. 2022], PaperEdge [Ma et al. 2022] and DocGeoNet [Feng et al. 2022].

Method	MS-SSIM \uparrow	LD \downarrow	AD \downarrow	CER \downarrow	ED \downarrow	Para.
DewarpNet	0.472	8.38	0.396	0.217	834	86.9M
DispFlow	0.432	7.62	0.396	0.292	1216	23.6M
DocTr	<i>0.509</i>	7.78	<u>0.366</u>	<i>0.181</i>	<u>712</u>	26.9M
PW Unwarping	0.490	8.65	0.431	0.252	987	-
DDCP	0.473	8.93	0.423	0.278	1118	13.3M
FDRNet	<u>0.543</u>	8.08	0.396	0.214	878	-
RDGR	0.495	8.50	0.432	0.171	732	-
Marior	0.476	<u>7.37</u>	0.404	0.200	797	-
PaperEdge	0.472	7.98	<i>0.367</i>	0.193	763	36.6M
DocGeoNet	0.504	7.70	0.378	0.190	736	24.8M
Ours	0.544	6.83	0.315	<u>0.172</u>	707	8M

Table 3: Quantitative unwarping performance comparisons on our UVDoc benchmark dataset. Bold font indicates best, underline indicates second-best and *italic* indicates third-best score. Refer to Table 2 for the list of referenced methods.

Method	MS-SSIM \uparrow	AD \downarrow	CER \downarrow	ED \downarrow	H-line \downarrow	V-line \downarrow
DewarpNet	0.600	0.189	0.115	338	3.22	4.32
DocTr	<i>0.684</i>	<i>0.176</i>	0.065	<u>192</u>	2.42	<u>3.32</u>
DDCP	0.591	0.334	0.117	362	4.20	4.88
RDGR	0.603	0.314	0.065	190	4.03	5.87
DocGeoNet	<u>0.714</u>	<u>0.167</u>	<u>0.066</u>	<i>196</i>	<u>2.24</u>	<i>3.91</i>
Ours	0.784	0.122	<i>0.072</i>	202	1.82	2.48

to horizontal (resp. vertical) lines; see Fig. 6 for a visual explanation of the process. Note that unlike other metrics that are commonly used in the document unwarping field to measure distortion (such as LD and AD), this metric does not rely on the usage of dense SIFT flow, which is slow to compute and can give unstable results due to shading artifacts.

6 EXPERIMENTS

6.1 Evaluation

We evaluate our network on the DocUNet benchmark dataset [Ma et al. 2018] as well as on our own UVDoc benchmark, described in Sec. 5.1. The DocUNet benchmark is composed of 65 documents. For each of them, 2 deformed images in a real-world scenario are provided. The ground truth flatbed-scans are also provided for comparison. Note that similarly to Feng et al. [2022] we exclude the two images of document 64, as the real world images are rotated by 180 degrees. We also exclude this document when computing the quantitative results for previous works.

Quantitative evaluation. We compare our method with several state-of-the-art deep learning methods. For each of them, we use the DocUNet result images published by the authors. We also evaluate the methods that additionally published their pre-trained models on our UVDoc benchmark. All metric scores are evaluated using Tesseract v4.0.0, pytesseract v0.3.10, MATLAB R2022a, Levenshtein v0.21.0 and jiwer v3.0.1. The results are presented in Tables 2 and 3.

Compared to previous works, our method achieves state-of-the-art MS-SSIM, LD, AD and ED performance and a second-best CER score on the DocUNet benchmark. On the UVDoc benchmark our method achieves state-of-the-art performance on visual metrics (MS-SSIM and LD). On OCR metrics, our network performs close to state-of-the-art. Small differences should be interpreted with care, as OCR scores suffer from high standard deviation (see Table 5). This is due to the Tesseract OCR engine being sensitive to small changes in input, as explained in its documentation [tesseract-ocr 2023]. Our method also ranks best on the horizontal and vertical line straightness metrics (Sec. 5.2.1), indicating better unwarping of the geometric features.

Our approach builds on a grid-based unwarping method, thanks to which our network is significantly smaller in size than current state-of-the-art methods, whilst still achieving state-of-the-art performance. We compare our network size to previous works in the last column of Table 2.

In addition to the performance of our own method, we also evaluate the effect of adding our UVDoc data to the DewarpNet [Das et al. 2019] architecture. We compare the performance of the pre-trained DewarpNet models fine-tuned for 10 epochs on the Doc3D data with the performance of the pre-trained models fine-tuned for 5 epochs on a combination of Doc3D and UVDoc data. As shown in Table 4, adding the UVDoc data into the fine-tuning process greatly improves all metrics except for the OCR performance on the *shaded* DocUNet images.

Qualitative evaluation. In addition to the quantitative comparisons made in the previous section, we provide a qualitative comparison to previous works. We show a side-by-side comparison of unwrapped images by several methods in Fig. 7 and Fig. 8. The left-most column shows the input images. The images unwrapped by our method are perceptually of high quality and have good unwarping at the borders of the document as well, even though we do not include explicit handling of borders or segmentation, in contrast to [Feng et al. 2021a, 2022; Ma et al. 2022; Zhang et al. 2022]. We present more qualitative results on real-world images in Fig. 1. We include the unwrapped images for all items in the DocUNet and UVDoc benchmarks in the supplemental material.

6.2 Ablation study

We show the effectiveness of the dual-task learning, i.e., the combination of predicting the 3D and the 2D grid meshes in the training process, the effectiveness of the reconstruction loss \mathcal{L}_r , as well as the benefit of combined training on both Doc3D and our UVDoc dataset, via ablation experiments. As we notice large variance in the OCR performance, we use averages of 10 repeated experiments with constant settings to perform the ablation study.

We first show in Table 5 that training on a combination of the Doc3D and UVDoc datasets considerably improves the performance

on all metrics, compared to training only on the Doc3D data. To ensure a fair comparison between the two, we double the number of epochs for the Doc3D-only training (both the number of epochs at a constant learning rate, as well as the number of epochs with linearly decaying learning rate), such that they process the same number of samples and have an equal amount of optimizer steps. In particular, adding the UVDoc data to the training process leads to improvements of 8.9% for MS-SSIM, 12.9% for LD and 9.7% for AD on the DocUNet benchmark. We attribute the improvement in performance to the fact that our data (UVDoc) is closer in appearance to the real document photographs in the DocUNet dataset, as well as the fact that the 3D ground-truth data in our dataset is more physically accurate (albeit coarser), since we do not apply any smoothing to it.

We also show that the dual-task of learning the 3D grid W along with the 2D grid G improves the performance by comparing it against our full model but with the loss on W removed. Table 5 shows that including the L_1 loss on W greatly improves all non-OCR metrics on both DocUNet and UVDoc benchmarks whilst the OCR metrics remain comparable.

Table 5 additionally shows the benefits of the reconstruction loss. Adding it improves performances on almost all metrics, and those that worsen only do so slightly. This loss helps the model to target content in the document, making the unwarping better in the areas that matter the most.

7 CONCLUSION

We presented UVDoc, a new document unwarping dataset that consists of pseudo-photorealistic images of warped documents along with annotated ground-truth 3D shapes and unwarping functions. Our proposed acquisition methodology is simple to implement and uses relatively inexpensive, commonly attainable equipment, enabling easy replication and further expansion of the dataset by others. Since our dataset includes both a shaded and unshaded version of each document image, it allows its users to evaluate unwarping performance without the influence of shading artifacts.

We show that with the addition of the UVDoc dataset, our dual-task deep learning approach that implicitly encodes the coupling between the document’s 3D shape and its appearance in a 2D photograph achieves state-of-the-art performance on the commonly used DocUNet benchmark. Additionally, we introduce the new UVDoc benchmark and a new line straightness metric, on which we also achieve state-of-the-art results.

The shape-from-shading effect can remain quite strong in the unwrapped documents and makes some of them appear more distorted to the human eye than they geometrically are. Research into the illumination correction process is therefore of great importance. Since the pseudo-photorealistic nature of our dataset allows us to decouple the deformation and shading of a warped document image, it could benefit research progress in this field.

ACKNOWLEDGMENTS

We thank Ruben Schenk for his help with data capture. This work was supported in part by the European Research Council (ERC) under the European Union’s Horizon 2020 Research and Innovation Programme (ERC Consolidator Grant, agreement No. 101003104, MYCLOTH).

Table 4: Quantitative comparison on the DocUNet and UVDoc benchmark datasets for DewarpNet [2019] finetuned with and without the UVDoc data. We finetuned the pre-trained models for 10 epochs with Doc3D only, and for 5 epochs with Doc3D+UVDoc to equalize the number of optimization steps.

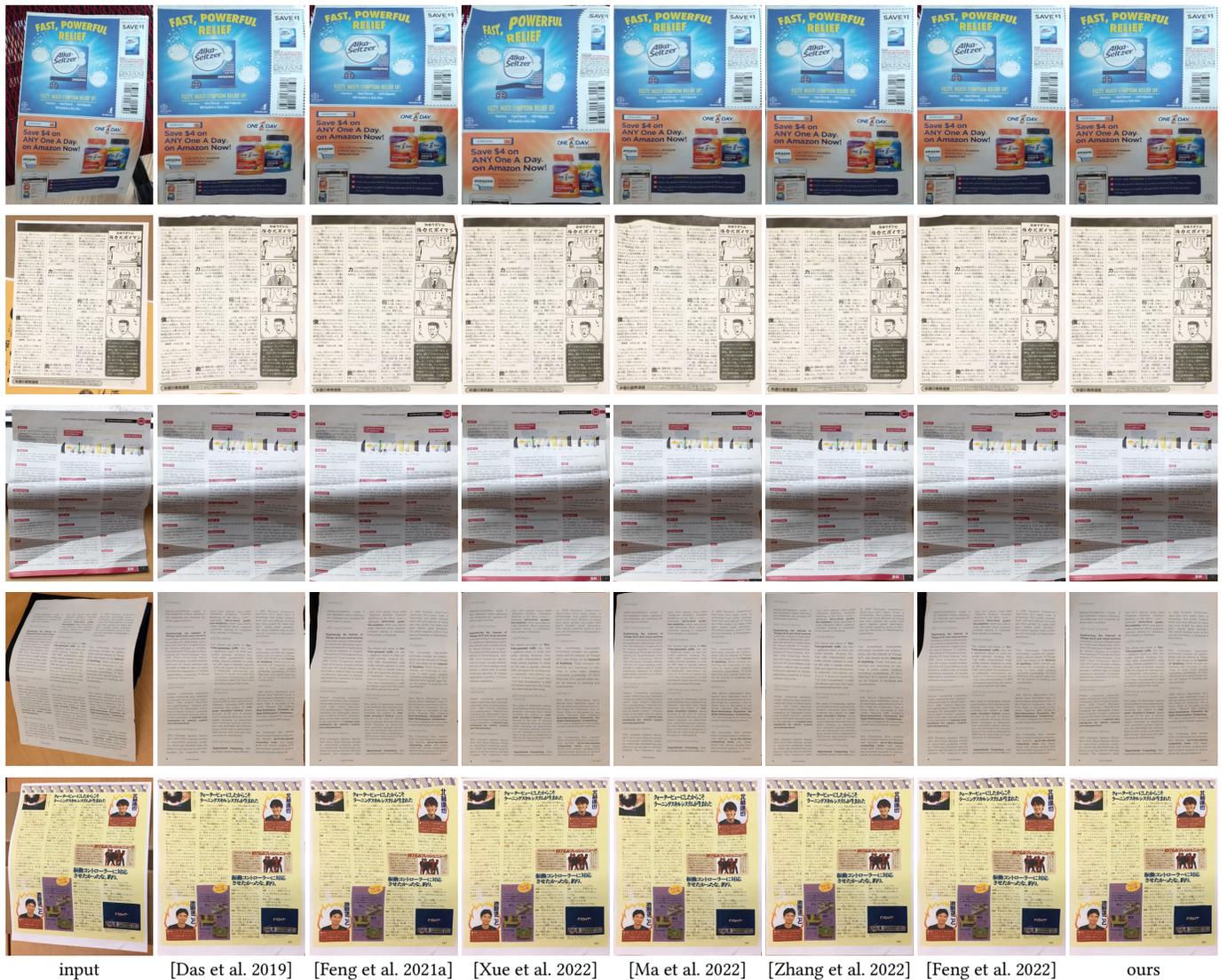
	DocUNet					UVDoc					
	MS-SSIM \uparrow	LD \downarrow	AD \downarrow	CER \downarrow	ED \downarrow	MS-SSIM \uparrow	AD \downarrow	CER \downarrow	ED \downarrow	H-line \downarrow	V-line \downarrow
Doc3D only	0.475	8.40	0.411	0.222	856	0.659	0.211	0.085	265	3.48	4.75
Doc3D + UVDoc	0.504	7.68	0.400	0.228	878	0.725	0.151	0.075	232	2.88	3.56

Table 5: Ablations about losses and data used. The reported values are averages and standard deviations over 10 repetitions of training with otherwise constant parameters on the DocUNet and UVDoc benchmark datasets. Settings used in our final model are underlined.

		DocUNet					UVDoc					
		MS-SSIM \uparrow	LD \downarrow	AD \downarrow	CER \downarrow	ED \downarrow	MS-SSIM \uparrow	AD \downarrow	CER \downarrow	ED \downarrow	H-line \downarrow	V-line \downarrow
<u>Doc3D + UVDoc</u>		0.536\pm0.006	6.96\pm0.17	0.325\pm0.006	0.195\pm0.012	745\pm34	0.762\pm0.014	0.129\pm0.008	0.070\pm0.010	205\pm23	1.85\pm0.06	2.53\pm0.06
Doc3D only		0.492 \pm 0.004	7.99 \pm 0.13	0.360 \pm 0.007	0.197 \pm 0.018	757 \pm 57	0.669 \pm 0.015	0.178 \pm 0.013	0.078 \pm 0.013	220 \pm 30	2.42 \pm 0.03	3.85 \pm 0.16
3D grid	w/	0.536\pm0.006	6.96\pm0.17	0.325\pm0.006	0.195 \pm 0.012	745 \pm 34	0.762\pm0.014	0.129\pm0.008	0.070 \pm 0.010	205 \pm 23	1.85\pm0.06	2.53\pm0.06
	w/o	0.531 \pm 0.005	7.04 \pm 0.16	0.331 \pm 0.004	0.189\pm0.017	743\pm54	0.747 \pm 0.010	0.148 \pm 0.011	0.068\pm0.010	201\pm22	1.87 \pm 0.08	2.59 \pm 0.08
\mathcal{L}_r	w/	0.536\pm0.006	6.96 \pm 0.17	0.325\pm0.006	0.195\pm0.012	745\pm34	0.762\pm0.014	0.129\pm0.008	0.070 \pm 0.010	205 \pm 23	1.85\pm0.06	2.53\pm0.06
	w/o	0.533 \pm 0.005	6.87\pm0.13	0.327 \pm 0.005	0.199 \pm 0.015	764 \pm 67	0.746 \pm 0.010	0.136 \pm 0.012	0.065\pm0.006	196\pm12	1.89 \pm 0.09	2.56 \pm 0.13

REFERENCES

- Stability AI. 2023. DeepFloyd IF. <https://github.com/deep-floyd/IF>.
- M.S. Brown and W.B. Seales. 2001. Document restoration using 3D shape: a general deskewing algorithm for arbitrarily warped documents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Vol. 2. 367–374 vol.2. <https://doi.org/10.1109/ICCV.2001.937649>
- M.S. Brown and W.B. Seales. 2004. Image restoration of arbitrarily warped documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26, 10 (2004), 1295–1306. <https://doi.org/10.1109/TPAMI.2004.87>
- Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing Textures in the Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, USA, 3606–3613. <https://doi.org/10.1109/CVPR.2014.461>
- Sagnik Das, Ke Ma, Zhixin Shu, and Dimitris Samaras. 2022. Learning an Isometric Surface Parameterization for Texture Unwrapping. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 580–597.
- Sagnik Das, Ke Ma, Zhixin Shu, Dimitris Samaras, and Roy Shilkrot. 2019. DewarpNet: Single-Image Document Unwarping With Stacked 3D and 2D Regression Networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 131–140. <https://doi.org/10.1109/ICCV.2019.00022>
- Sagnik Das, Kunwar Yashraj Singh, Jon Wu, Erhan Bas, Vijay Mahadevan, Rahul Bhotika, and Dimitris Samaras. 2021. End-to-end Piece-wise Unwarping of Document Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 4268–4277. <https://doi.org/10.1109/ICCV48922.2021.00423>
- Hao Feng, Shaokai Liu, Jiajun Deng, Wengang Zhou, and Houqiang Li. 2023. Deep Unrestricted Document Image Rectification. <https://doi.org/10.48550/arXiv.2304.08796>
- Hao Feng, Yuechen Wang, Wengang Zhou, Jiajun Deng, and Houqiang Li. 2021a. DocTr: Document Image Transformer for Geometric Unwarping and Illumination Correction. In *Proceedings of the ACM International Conference on Multimedia*. 273–281. <https://doi.org/10.1145/3474085.3475388>
- Hao Feng, Wengang Zhou, Jiajun Deng, Qi Tian, and Houqiang Li. 2021b. DocScanner: Robust Document Image Rectification with Progressive Learning. <https://doi.org/10.48550/arXiv.2110.14968>
- Hao Feng, Wengang Zhou, Jiajun Deng, Yuechen Wang, and Houqiang Li. 2022. Geometric Representation Learning for Document Image Rectification. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 475–492. https://doi.org/10.1007/978-3-031-19836-6_27
- Felix Hertlein, Alexander Naumann, and Patrick Philipp. 2023. Inv3D: a high-resolution 3D invoice dataset for template-guided single-image document unwarping. *International Journal on Document Analysis and Recognition (IJ DAR)* (29 Apr 2023), 12 pages. <https://doi.org/10.1007/s10032-023-00434-x>
- Xiangwei Jiang, Rujiao Long, Nan Xue, Zhibo Yang, Cong Yao, and Guisong Xia. 2022. Revisiting Document Image Dewarping by Grid Regularization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4533–4542. <https://doi.org/10.1109/CVPR52688.2022.00450>
- TaeHo Kil, Wonkyo Seo, Hyung Il Koo, and Nam Ik Cho. 2017. Robust Document Image Dewarping Method Using Text-Lines and Line Segments. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*. 865–870. <https://doi.org/10.1109/ICDAR.2017.146>
- Beom Su Kim, Hyung Il Koo, and Nam Ik Cho. 2015. Document dewarping via text-line based optimization. *Pattern Recognition* 48, 11 (2015), 3600–3614. <https://doi.org/10.1016/j.patcog.2015.04.026>
- Hyung Il Koo, Jinho Kim, and Nam Ik Cho. 2009. Composition of a Dewarped and Enhanced Document Image From Two View Images. *IEEE Transactions on Image Processing* 18, 7 (2009), 1551–1562. <https://doi.org/10.1109/TIP.2009.2019301>
- Xiaoyu Li, Bo Zhang, Jing Liao, and Pedro V. Sander. 2019. Document rectification and illumination correction using a patch-based CNN. *ACM Transactions on Graphics (TOG)* 38 (2019), 1–11.
- Jian Liang, D. DeMenthon, and D. Doermann. 2005. Flattening curved documents in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 2. 338–345 vol. 2. <https://doi.org/10.1109/CVPR.2005.163>
- Jian Liang, Daniel DeMenthon, and David S. Doermann. 2008. Geometric Rectification of Camera-Captured Document Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30 (2008), 591–605. <https://doi.org/10.1109/TPAMI.2007.70724>
- Xiyuan Liu, Gaofeng Meng, Bin Fan, Shiming Xiang, and Chunhong Pan. 2020. Geometric rectification of document images using adversarial gated unwarping network. *Pattern Recognition* 108 (2020), 107576. <https://doi.org/10.1016/j.patcog.2020.107576>
- Dong Luo and Pengbo Bo. 2022. Geometric Rectification of Creased Document Images based on Isometric Mapping. <https://doi.org/10.48550/arXiv.2212.08365>
- Ke Ma, Sagnik Das, Zhixin Shu, and Dimitris Samaras. 2022. Learning From Documents in the Wild to Improve Document Unwarping. In *Proceedings of ACM SIGGRAPH*. 9 pages. <https://doi.org/10.1145/3528233.3530756>
- Ke Ma, Zhixin Shu, Xue Bai, Jue Wang, and Dimitris Samaras. 2018. DocUNet: Document Image Unwarping via a Stacked U-Net. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4700–4709. <https://doi.org/10.1109/CVPR.2018.00494>
- Gaofeng Meng, Yuanqi Su, Ying Wu, Shiming Xiang, and Chunhong Pan. 2018. Exploiting Vector Fields for Geometric Rectification of Distorted Document Images. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 180–195. https://doi.org/10.1007/978-3-030-01270-0_11
- Gaofeng Meng, Ying Wang, Shenquan Qu, Shiming Xiang, and Chunhong Pan. 2014. Active Flattening of Curved Document Images via Two Structured Beams. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3890–3897. <https://doi.org/10.1109/CVPR.2014.497>
- C. Nachappa, N Shobha Rani, Peeta Basa Pati, and M. Gokulnath. 2023. Adaptive dewarping of severely warped camera-captured document images based on document map generation. *International Journal on Document Analysis and Recognition (IJ DAR)* 26 (2023), 149–169. <https://doi.org/10.1007/s10032-022-00425-4>
- tesseract-ocr. 2023. *Tesseract user manual*. <https://tesseract-ocr.github.io/tessdoc/ImproveQuality.html>
- Yuandong Tian and Srinivasa G Narasimhan. 2011. Rectification and 3D reconstruction of curved document images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 377–384. <https://doi.org/10.1109/CVPR.2011.5995540>
- Yau-Chat Tsoi and Michael S. Brown. 2007. Multi-View Document Rectification using Boundary. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1–8. <https://doi.org/10.1109/CVPR.2007.383251>
- Adrian Ulges, Christoph H. Lampert, and Thomas Breuel. 2004. Document Capture Using Stereo Vision. In *Proceedings of the 2004 ACM Symposium on Document Engineering*. 198–200. <https://doi.org/10.1145/1030397.1030434>
- Guo-Wang Xie, Fei Yin, Xu-Yao Zhang, and Cheng-Lin Liu. 2020. Dewarping Document Image by Displacement Flow Estimation with Fully Convolutional Network. In *International Workshop on Document Analysis System*. 131–144.
- Guo-Wang Xie, Fei Yin, Xu-Yao Zhang, and Cheng-Lin Liu. 2021. Document Dewarping with Control Points. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*. 466–480. https://doi.org/10.1007/978-3-030-86549-8_30
- Zhen Xu, Fei Yin, Peipei Yang, and Cheng-Lin Liu. 2022. Document Image Rectification in Complex Scene Using Stacked Siamese Networks. In *Proceedings of the 26th International Conference on Pattern Recognition (ICPR)*. 1550–1556. <https://doi.org/10.1109/ICPR56361.2022.9956331>
- C. Xue, Z. Tian, F. Zhan, S. Lu, and S. Bai. 2022. Fourier Document Restoration for Robust Document Dewarping and Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4563–4572. <https://doi.org/10.1109/CVPR52688.2022.00453>
- A. Yamashita, A. Kawarago, T. Kaneko, and K.T. Miura. 2004. Shape reconstruction and image restoration for non-flat surfaces of documents with a stereo vision system. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, Vol. 1. 482–485. <https://doi.org/10.1109/ICPR.2004.1334171>
- Shaodi You, Yasuyuki Matsushita, Sudipta Sinha, Yusuke Bou, and Katsushi Ikeuchi. 2018. Multiview Rectification of Folded Documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 2 (2018), 505–511. <https://doi.org/10.1109/TPAMI.2017.2675980>
- Jiaxin Zhang, Canjie Luo, Lianwen Jin, Fengjun Guo, and Kai Ding. 2022. Marior: Margin Removal and Iterative Content Rectification for Document Dewarping in the Wild. In *Proceedings of the ACM International Conference on Multimedia*. 2805–2815. <https://doi.org/10.1145/3503161.3548214>
- L. Zhang and C.L. Tan. 2005. Warped image restoration with applications to digital libraries. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*. 192–196. <https://doi.org/10.1109/ICDAR.2005.252>
- Li Zhang, Yu Zhang, and Chew Tan. 2008. An Improved Physically-Based Method for Geometric Restoration of Distorted Document Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 4 (2008), 728–734. <https://doi.org/10.1109/TPAMI.2007.70831>
- Zheng Zhang, Chew Lim Tan, and Liying Fan. 2004. Restoration of curved document images through 3D shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1. 1–I. <https://doi.org/10.1109/CVPR.2004.1315007>



input

[Das et al. 2019]

[Feng et al. 2021a]

[Xue et al. 2022]

[Ma et al. 2022]

[Zhang et al. 2022]

[Feng et al. 2022]

ours

Figure 7: Qualitative comparisons on the DocUNet benchmark dataset. From left to right: input, DewarpNet [Das et al. 2019], DocTr [Feng et al. 2021a], FDRNet [Xue et al. 2022], PaperEdge [Ma et al. 2022], Marior [Zhang et al. 2022], DocGeoNet [Feng et al. 2022], ours. All input images come from the “crop” subset.

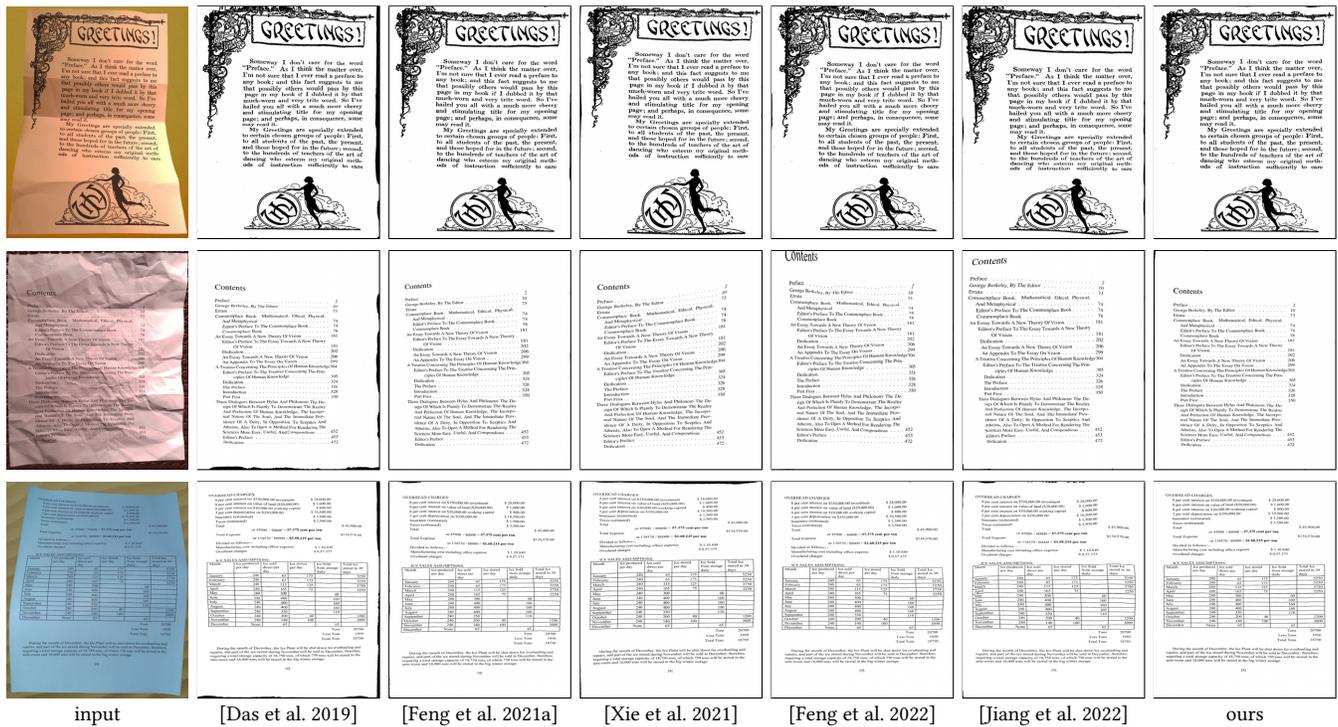


Figure 8: Qualitative comparisons on the UVDoc benchmark dataset. From left to right: input, DewarpNet [Das et al. 2019], DocTr [Feng et al. 2021a], DDCP [Xie et al. 2021], DocGeoNet [Feng et al. 2022], RDGR [Jiang et al. 2022], ours. All input images come from the shaded subset and we show the unshaded version of the unwarping images to emphasize their structure.