Learning From Documents in the Wild to Improve Document Unwarping

Ke Ma
Snap Inc.
New York, NY, USA
kemma@cs.stonybrook.edu

Sagnik Das
Stony Brook University
Stony Brook, NY, USA
sadas@cs.stonybrook.edu

Zhixin Shu
Adobe Research
San Jose, CA, USA
zshu@adobe.com

Dimitris Samaras
Stony Brook University
Stony Brook, NY, USA
samaras@cs.stonybrook.edu

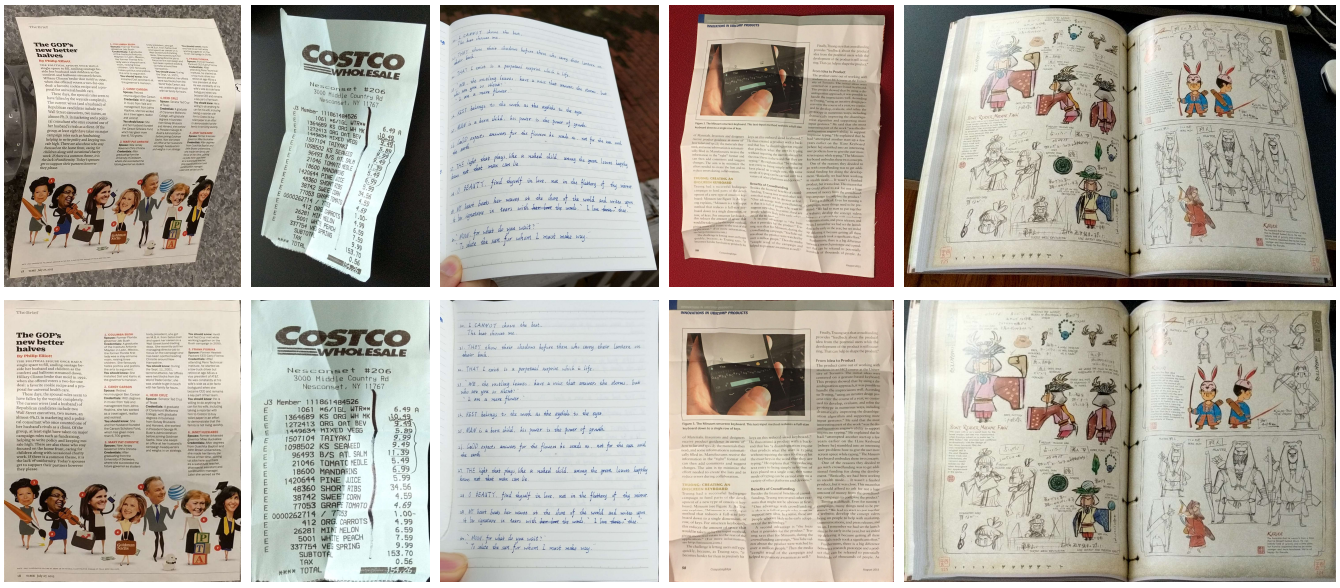

Figure 1: PaperEdge unwarping results. The top row is the casual photos of documents. The bottom row is the unwarping results by our proposed model: PaperEdge. PaperEdge utilizes both synthetic and real document images in training. This model is excellent at unwarping documents in different deformation and content. The unwarped documents from PaperEdge achieves 16.2% **Word Error Rate reduction in the OCR task, compared to the results from the previous unwarping method [Das et al. 2019].**

ABSTRACT

Document image unwarping is important for document digitization and analysis. The state-of-the-art approach relies on purely synthetic data to train deep networks for unwarping. As a result, the trained networks have generalization limitations when testing on real-world images, often yielding unsatisfying results. In this work, we propose to improve document unwarping performance by incorporating real-world images in training. We collected Document-in-the-Wild (DIW) dataset contains 5000 captured document images with large diversities in content, shape, and capturing environment. We annotate the boundaries of all DIW images and use them for weakly supervised learning. We propose a novel network architecture, *PaperEdge*, to train with a hybrid of synthetic and real document images. Additionally, we identify and analyze the flaws of popular evaluation metrics, e.g., MS-SSIM and Local Distortion (LD), for document unwarping and propose a more robust and reliable error metric called Aligned Distortion (AD). Training with a combination of synthetic and real-world document images, we demonstrate state-of-the-art performance on popular benchmarks with comprehensive quantitative evaluations and ablation studies. Code and data are available at https://github.com/cvlab-stonybrook/PaperEdge.

SIGGRAPH '22 Conference Proceedings, August 7–11, 2022, Vancouver, BC, Canada

ACM ISBN 978-1-4503-9337-9/22/08…$15.00
https://doi.org/10.1145/3528233.3530756

CCS CONCEPTS

• **Applied computing** → **Document capture**; • **Computing methodologies** → **Computer vision**.

KEYWORDS

document image unwarping, datasets, convolutional neural networks

ACM Reference Format:

Ke Ma, Sagnik Das, Zhixin Shu, and Dimitris Samaras. 2022. Learning From Documents in the Wild to Improve Document Unwarping. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings (SIGGRAPH '22 Conference Proceedings), August 7–11, 2022, Vancouver, BC, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3528233.3530756

1 INTRODUCTION

Compared to traditional paper documents, digital documents are much easier to archive, edit, sign, and share. Nowadays, more and more physical documents are digitized for efficient workflows. During the COVID-19 pandemic, digital documents also play a vital role for business as physical contact is restricted. Ubiquitous smartphones, equipped with high-quality cameras, have made photographing a document the standard way to digitize it. However, documents in these photos are always distorted due to uncontrolled paper geometry and capturing condition. These distortions hamper information extraction from these documents, decrease readability, and break downstream automatic document analysis pipelines such as layout extraction and Optical Character Recognition (OCR), which were built to process only document scans.

Both model-driven and data-driven methods have been proposed to address document rectification. Model-driven methods utilize explicit geometric models to fit the deformed document surface. They usually involve slow optimization steps to obtain unwarped results, which are unsuitable for real-time applications. Recently, data-driven methods have gained popularity. These methods train an unwarping neural network to map a casual document image to a deformation field that warps the deformed input into a rectified, scan-like result. Such networks can achieve real-time performance.

The training data plays a central role in building a well generalizable unwarping network in the data-driven setting. Ideally, one would want to collect sufficient real-world training data for supervised learning: captured images of distorted documents and ground truth deformations (usually represented by some image warping functions). However, this type of data is hard to obtain due to difficulties in large-scale accurate 3D reconstruction and dense registration in the wild. Previous work [Das et al. 2019; Ma et al. 2018] resorted to training on synthetic document images with ground truth warping fields. Existing data synthesis schemes deform a flat document with a known warping field. A network is then trained to regress the warping field from the deformed image, which is subsequently used to "unwarp" the image into a flat document. However, synthesizing hyper-realistic warped, creased, and crumpled document paper is very challenging due to the complexity in modeling geometry and material. Images from state-of-the-art synthetic document dataset [Das et al. 2019] are visibly different from real-world images. In fact, we also demonstrate in the supplementary material, there is data redundancy in the synthetic dataset. More specifically, after training with 32,000 synthetic images (32% of the entire dataset), the performance improvement from additional synthetic training data becomes insignificant.

We propose to improve document unwarping by introducing PaperEdge, the first unwarping model trainable with real-world document images. It is non-trivial to incorporate real images in the prior supervised learning approaches [Li et al. 2019; Markovitz et al. 2020] due to the absence of ground truth deformation, which is difficult to obtain for real-world documents. PaperEdge enables learning from both synthetic and real-world training images: For synthetic data, we train in a supervised manner using ground truth deformation. For real-world images without ground truth deformation, we utilize the document edges [Gumerov et al. 2004; Tsoi and Brown 2007] as weak supervision. Document edges reflect a global rectangular shape deformation; therefore, can be used as a training signal. They are also straightforward to annotate with off-the-shelf image segmentation tools [Rother et al. 2004]. To facilitate the proposed training scheme, we collected the Documents In the Wild (DIW) dataset with 5000 document photos and their edge annotations.

Moreover, we introduce a texture-based warping model to further enhance the results. Document image texture provides valuable cues for unwarping because document content is usually structured. As edges are effective for global image unwarping, the texture is beneficial for recovering local distortions. We propose a self-supervised learning strategy [Gidaris et al. 2018; Zhang et al. 2019] to train the texture-aware component. In practice, we augment each training sample with a randomly generated deformation perturbation to form a training image pair. After that, we train the network in a siamese style [Koch et al. 2015] using these pairs.

We also demonstrate that popular quantitative evaluation criteria such as MS-SSIM and Local Distortion (LD) are not ideal for evaluating document image unwarping. We show that (1) MS-SSIM is very sensitive to perceptually negligible perturbations, and (2) LD calculation accounts for a large amount of unimportant error at textureless regions. To address this issue, we introduce Aligned Distortion (AD), a more robust quantitative measure for evaluating document unwarping performance.

We summarize our contributions as follows: (1) we propose a novel network architecture for learning document unwarping. It is the first method that can be trained with both synthetic document images and camera-captured casual document images; (2) we propose Aligned Distortion (AD) – a robust evaluation metric for document unwarping; (3) we contribute a new document image dataset with 5000 in-the-wild document images and their edge annotations; (4) we achieve state-of-the-art performance on the benchmark [Ma et al. 2018] under all evaluation criterion.

2 PREVIOUS WORK

Document unwarping has been extensively studied in the literature. We roughly categorize prior work into model-driven methods and data-driven methods.

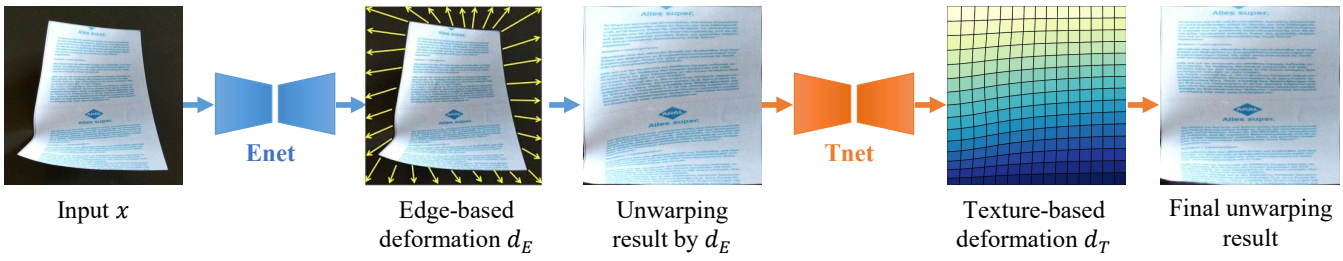

Figure 2: PaperEdge system pipeline. Our system consists of two subnetworks: Enet and Tnet. Enet provides a coarse, document edge-based unwarping that rectifies the overall shape; Tnet further enhance the result with local deformation learned from document texture.

2.1 Model-driven methods

Model-driven methods usually consist of two-steps: (1) estimate document surface deformation and (2) flatten the deformed surface.

Surface Deformation Estimation. [Meng et al. 2014] used laser beams to estimate a developable surface. [Courteille et al. 2007] applied Shape from Shading [Wada et al. 1997] on documents. Both [Ulges et al. 2004] and [You et al. 2017] utilized multiple images to reconstruct the 3D shape. Shape from Template can also be used for estimate surface shape as shown by [Bartoli and Collins 2013; Chhatkuli et al. 2014; Khan et al. 2014]. For document images, several methods [Ezaki et al. 2005; Kil et al. 2017; Liang et al. 2008; Liu et al. 2015; Lu and Tan 2006; Meng et al. 2018; Ulges et al. 2004] have been proposed to estimate shape from document components such as text lines, blocks, and figures, *etc.*

Surface Flattening. [Kim et al. 2015] and [Kil et al. 2017] modeled the surface as a Generalized Cylindrical Surface (GCS) and demonstrated that the deformation is invertible in parameter space. [Liang et al. 2008] and [Meng et al. 2015] both approximated the surface with finite number of planar strips on the tangent planes. [Tian and Narasimhan 2011], and [Meng et al. 2018] built a sparse correspondence between a projected mesh on the input image and a imaginary flattened image to recover the texture. [You et al. 2017] used conformal mapping to flattened the mesh.

2.2 Data-driven methods

Recently, deep neural network have been widely adopted to learn paper unwarping and rectification. [Shafait and Breuel 2007] released a dataset of 102 binarized images. [Pumarola et al. 2018] utilized a CNN to estimate the vertex coordinate on a regular mesh grid for a deformed surface. [Jiménez et al. 2018] embedded the SfT framework in a CNN. [Das et al. 2017] trained a CNN to detect the folding edges. [Ma et al. 2018] proposed an end-to-end unwarping network trained on randomly perturbed 2D document images. [Li et al. 2019] extended this idea with a local/global two-branch network. [Liu et al. 2020] utilized gated network blocks and an adversarial loss to improve the results. [Xie et al. 2020] estimated the deformation offset instead of the absolute deformation field and incorporated a local smooth constraint. [Das et al. 2019] proposed DewarpNet to explicitly modeling the deformed 3D shape and introduced the Doc3D dataset with about 100K rendered images. [Markovitz et al. 2020] followed similar data synthesis pipeline and further augmented DewarpNet with text block angle supervision. [Das et al. 2021] proposed a patch-wise approach for better local unwarping. More recently, [Feng et al. 2021] introduced Transformers [Vaswani et al. 2017] as a stronger backbone.

State-of-the-art synthetic datasets such as Doc3D provide ground truth deformations for supervised training. However, the visual discrepancy between synthetic data and real data is very significant. Unlike prior work, our framework can utilize real data for training with inexpensive paper edge annotations, which significantly improves the network's generalization ability. As a result, we achieve state-of-the-art performance on unwarping casual document images by utilizing both synthetic and real images. We noticed a concurrent study [Xue et al. 2022] that also exploited real world images to improve unwarping. It required document scans as annotations but scanners are not always available. While we propose to use more accessible document boundary annotations.

3 METHOD

Our system PaperEdge contains two sub-networks that unwarp an input document image in two steps (Fig. 2.): The first sub-network "Enet" unwarps the input using document edge information. The output of Enet is a warping field for a coarse scale, "global" deformation that warps the input document image into a shape with desired boundary property, i.e., a rectangle. The second sub-network, "Tnet" outputs a fine-scale, "local" warping field relying on document texture. It corrects the local deformation of the previous output, straightens the text line, and rectifies the content shape.

Formally, given a deformed document image x, the unwarping result x_t is obtained by

$$x_t = \phi(\phi(x, d_E), d_T), \tag{1}$$

where $\phi(a_1, a_2)$ is a 2D warping function that warps a_1 based on a deformation field a_2. Note that a_1 can be an image or another deformation field [Ashburner 2007]. In Eq. 1, d_E is the edge-based deformation field outputted by Enet, and d_T is the texture-based deformation field outputted by Tnet. A deformation field defines a backward mapping [Chen et al. 1999], which determines the position in the input image to be sampled and mapped to the target.

In this section, we describe Enet and Tnet, including their network architectures and training methods. The training methods of using synthetic and real-world data are different as they provide different training signals. We will describe them separately and

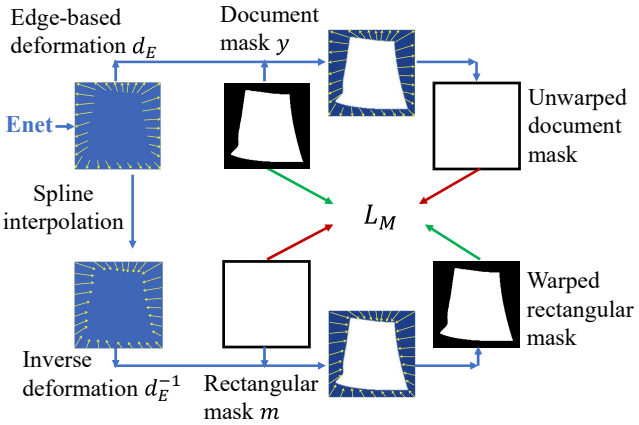

Figure 3: Cycle consistent mask loss in Enet. We infer d_E^{-1} by differentiable spline interpolation from d_E. The forward cycle loss is to warp the ground truth document mask y with d_E to match the rectangular mask m (red arrows). The backward cycle loss is to warp m with d_E^{-1} to match the ground truth mask y (green arrows).

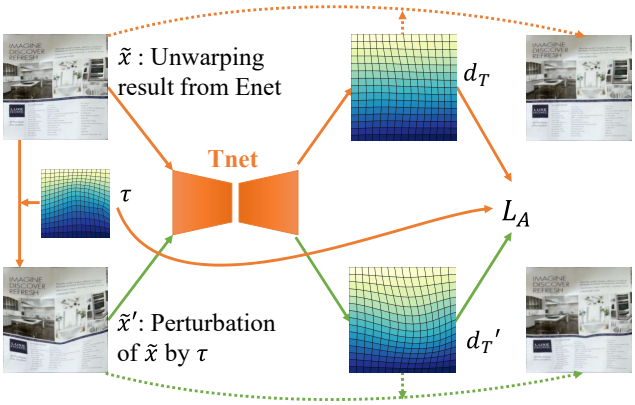

Figure 4: Self-supervised training for Tnet. $\tilde{x}$ is the output from Enet. We warp $\tilde{x}$ with a known deformation perturbation τ, obtaining $\tilde{x}'$. Tnet is trained in a siamese network manner. Though d_T and d_T' are unknown, they are associated by τ in L_S.

show that our network design allows hybrid training with both data types.

3.1 Enet: Edge-based Unwarping

Enet is a fully-convolutional encoder-decoder. The encoder has 6 residual blocks [He et al. 2016] and each block down-samples the input feature map by a factor of 2. The decoder has 4 residual blocks and each block up-samples the input feature map by a factor of 2. For all our experiments, we used $256 \times 256 \times 5$ (RGB of input image x + coordinates [Liu et al. 2018]) as Enet input and the bottleneck feature map is 4×4×512. The deformation field (backward mapping) output from the decoder is $64 \times 64 \times 2$.

Supervised training on synthetic images. Given a synthetic deformed document image x and its ground truth deformation field d^*, we train Enet in a fully supervised manner. Particularly, assuming $d^* \in \mathbb{R}^{N \times N \times 2}$ where N is the spatial resolution, the boundary elements of d^*: $\{d^*(i,1), d^*(i,N), d^*(1,j), d^*(N,j) \mid i,j \in [1,N]\}$ indicate the coordinates of document boundaries in the input image. d_E has the same size as d^*. We train Enet to match the boundary elements of d_E and d^* with the following loss:

$$L_{SE} = |B(d_E) - B(d^*)|_1. \tag{2}$$

where B is a function extracting the boundary of a deformation field. These boundary elements are sufficient to infer the edge-based deformation field: the interior of d_E: $\{d_E(i,j) \mid i,j \in [2, N-1]\}$ is linearly interpolated from the boundaries.

Weakly supervised training on real images. Given a real-world warped document image x with its segmentation mask y, which is an easy-to-obtain edge map representation, we trained Enet in a weakly supervised manner. A loss on the segmentation mask is defined as:

$$L_M = |\phi(y, d_E) - m|_1, \tag{3}$$

where d_E is the output from Enet. m is a rectangular mask. Ideally, a successfully trained d_E can rectify the document image x and warp the mask y to a rectangle m. However, there is a trivial but wrong solution to this training loss: d_E could be a simple scale-up transformation that enlarges the mask region in y to fill the whole image area and result in the sought rectangular mask m. Inspired by ProAlignNet [Veeravasarapu et al. 2020], we introduce a cycle-consistent segmentation mask loss to address this problem:

$$L_M = |\phi(y, d_E) - m|_1 + |\phi(m, d_E^{-1}) - y|_1, \tag{4}$$

where d_E^{-1} is the inverse deformation field of d_E. Both d_E and d_E^{-1} define a mapping between two sets of pixel coordinates: $\mathbb{P} = \{\mathbf{p}_i\}$ from the source image and $\mathbb{Q} = \{\mathbf{q}_i\}$ from the target. The domain of d_E is the target image while the domain of d_E^{-1} is the source image. In other words, d_E is a backward mapping and d_E^{-1} a forward mapping. Thus d_E^{-1} defines a function $f(\mathbf{p}_i) = \mathbf{q}_i$.

In order to approximate d_E^{-1} from d_E, we use polyharmonic splines to fit this function:

$$f(\mathbf{p}) = \sum_i^n w_i \varphi(|\mathbf{p} - \mathbf{q}_i|_2) + \mathbf{v}^T \mathbf{p} + \mathbf{b}, \tag{5}$$

where $\varphi(r) = r$ is the polyharmonic spline radial basis function. n is the number of the correspondences on the document edges. Parameters w_i, $\mathbf{v}$, and $\mathbf{b}$ are obtained by minimizing $\sum_i^n |f(\mathbf{p}_i) - \mathbf{q}_i|$ which has a closed form solution. Then $d_E^{-1} = f(\mathbf{p}_i), \forall i$ in the source image. This cycle-consistent mask loss is illustrated in Fig. 3.

3.2 Tnet: Texture-based Unwarping

Tnet and Enet share the same network architecture but not weights. The input to Tnet is the unwarped image using the output of Enet $\tilde{x} = \phi(x, d_E)$. The output of Tnet is a deformation field $d_T \in \mathbb{R}^{64 \times 64 \times 2}$. After a step of coarse edge-based warping enabled by Enet, Tnet further improves the warping results by utilizing texture information for training.

Supervised training on synthetic images. With synthetic data, we obtain the ground truth texture-based deformation field d_T^* by separating d_E from ground truth deformation d^*. We utilize the associative property of the warping function [Ashburner 2007; Lin and Lucey 2017]: $\phi(\phi(a,b),c) = \phi(a,\phi(b,c))$. Per Eq. 1, we can derive $\phi(x,d^*) = \phi(\phi(x,d_E),d_T^*) = \phi(x,\phi(d_E,d_T^*))$. Hence $d^* = \phi(d_E,d_T^*)$ and $d_T^* = \phi(d_E^{-1},d^*)$ where d_E^{-1} is obtained in Eq. 4 and 5. The loss function is

$$L_{ST} = |d_T - d_T^*|_1. \tag{6}$$

Self-supervised training on real images. While training with real-world images where ground truth is not available, we train Tnet in a self-supervised manner. Our method is partially inspired by the Auto-Encoding Transformation (AET) [Zhang et al. 2019]. Given an input image $\tilde{x}$, we apply a random deformation perturbation τ to $\tilde{x}$ to obtain a perturbed image $\tilde{x}'$. Denoting the deformation field that unwarps $\tilde{x}'$ as d_T', the ground truth of neither d_T nor d_T' is available. However, $\tilde{x}$ and $\tilde{x}'$ must be unwarped to the same image as they were warped from the same original one. Thus we can train Tnet by minimizing:

$$\begin{aligned} L_A =& |\phi(\tilde{x},d_T) - \phi(\tilde{x}',d_T')|_1 \\ =& |\phi(\tilde{x},d_T) - \phi(\phi(\tilde{x},\tau),d_T')|_1. \end{aligned} \tag{7}$$

However, directly optimizing Eq. 7 yields poor results due to local minima. Therefore we simplify the above equation using the relation $\phi(\phi(\tilde{x},\tau),d_T') = \phi(\tilde{x},\phi(\tau,d_T'))$ Thus Eq. 7 is rewritten as:

$$L_A = |\phi(\tilde{x},d_T) - \phi(\tilde{x},\phi(\tau,d_T'))|_1, \forall \tilde{x}. \tag{8}$$

Because Eq. 8 holds for all $\tilde{x}$, optimizing the above function is equivalent to optimizing

$$L_A = |d_T - \phi(\tau,d_T')|_1, \tag{9}$$

as shown in Fig. 4. Similar to Eq. 4, we add a cycle consistency loss to Eq. 9. The final self-supervised learning objective is:

$$L_A = |d_T,\phi(\tau,d_T')|_1 + |\phi(\tau^{-1},d_T),d_T')|_1. \tag{10}$$

τ is constructed offline. We randomly select a point within the document area and move it in a random direction by a random offset as long as it is still within the document area. The deformation field for all the points is then interpolated by polyharmonic splines, which is similar to Eq. 5. This process is repeated a random number of $n \in [1,5]$ times to obtain τ. The inverse deformation field τ^{-1} is easy to obtain from τ following the same method that computes d_E^{-1} from d_E in Eq. 4.

4 DOCUMENTS-IN-THE-WILD DATASET

To demonstrate the effectiveness of real-world images, we built the Documents In the Wild (DIW) dataset. Physical documents are ubiquitous. The DIW dataset contains 5000 photos of about 600 daily-life documents, including 300 receipts, 10 books, 200 document sheets (academic papers, magazines, and advertising fliers), 50 product labels (ingredient/nutrition labels, clothes wash labels), etc. The annotation tool for edge annotations is created based on GrabCut [Rother et al. 2004]. On average, it only takes around 5 seconds to annotate one image. The final image is cropped based on the mask and resized to 512×512.

Table 1: The number of images in each document type in DIW.

Type	Receipt	Book	Sheet	Label
Number of images	2360	501	1823	316

Figure 5: DIW dataset: Sample images and the annotated document masks in the DIW dataset.

During data collection, we also annotate their types with one of the following labels: "receipt", "book", "sheet", and "label". We collected this meta information because we observed the document types provide useful prior about deformation. For example, the most common deformation on book pages is a simple curl, while receipts are often folded or crumpled. Thus, in the future, we could potentially model the deformation conditioned on document types. This prior is not used in this paper, but we believe this information will benefit future research. The number of each document type in the DIW dataset is shown in table 1. Sample images and corresponding document edge annotations (as masks) are shown in Fig. 5.

5 EXPERIMENTS

We train PaperEdge using our DIW dataset and synthetic dataset Doc3D [Das et al. 2019]. Doc3D contains 100,000 deformed document images and the corresponding annotations such as albedo, depth, and ground truth deformation *etc.* Following DewarpNet [Das et al. 2019], we use 88,000 labeled instances for training and the rest for validation. In DIW, all 5,000 images and the corresponding document mask annotations are used.

We evaluate PaperEdge on a popular benchmark dataset [Ma et al. 2018] with 130 in-the-wild document images of various deformations. In addition to the popular image similarity evaluation metrics such as MS-SSIM and Local Distortion (LD), we propose a more robust and reliable error measure for document unwarping called Aligned Distortion (AD). In this section, we first analyze the drawbacks of previous metrics and introduce our robust metric AD. We then compare our method with previous work using all evaluation metrics as well as OCR performance.

5.1 AD: A Robust Evaluation Metric

To quantitatively evaluate unwarping methods, Ma *et al.* [Ma et al. 2018] and Das *et al.* [Das et al. 2019] relied on MS-SSIM and LD to measure image similarity between unwarping results and ground-truth flat document images. However, we observe that these two metrics can sometimes be unreliable when measuring deformation errors (Fig. 6). The shift and scale sensitivity of MS-SSIM was previously noticed in [Markovitz et al. 2020]. In Fig. 6, we observe that

Table 2: Unwarping performance on benchmark. AD is the proposed image similarity metric. CER, WER and ED are OCR metrics.

Method		AD ↓	CER ↓	WER ↓	ED ↓	MS-SSIM ↑	LD ↓
Model-driven methods	Tian *et al.* [Tian and Narasimhan 2011]	2.112	0.675	0.797	2911	0.130	33.69
	Kim *et al.* [Kim et al. 2015]	0.903	0.287	0.412	1293	0.348	19.51
	Kil *et al.* [Kil et al. 2017]	0.679	0.270	0.391	1205	0.401	12.84
Data-driven methods	Ma *et al.* [Ma et al. 2018]	0.700	0.271	0.451	1197	0.439	10.90
	Li *et al.* [Li et al. 2019]	0.738	0.297	0.417	1289	0.383	12.83
	Das *et al.* [Das et al. 2019]	0.426	0.257	0.376	1131	**0.469**	8.98
	PaperEdge$_D$	**0.416**	**0.252**	**0.363**	**1108**	0.467	**8.79**
	PaperEdge$_{DD}$	**0.392**	**0.221**	**0.315**	**1010**	**0.470**	**8.50**

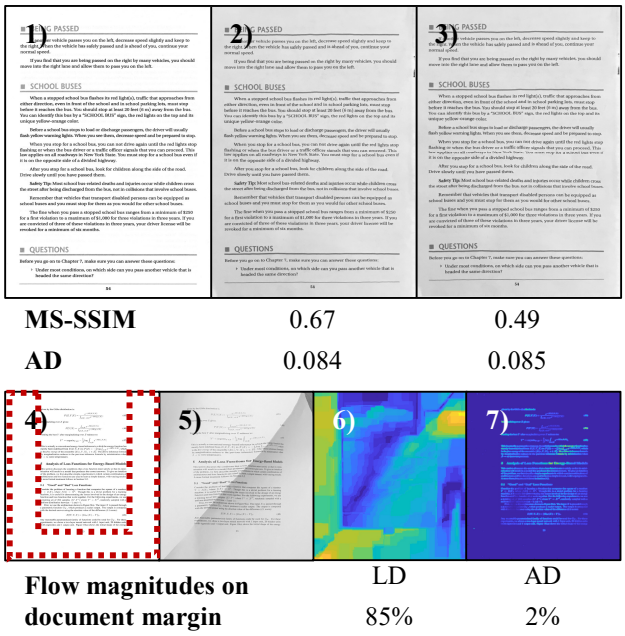

Figure 6: MS-SSIM and LD failures. 1) and 4) are document scans. 2) and 5) are unwarped results. 3) is 2) translated 10 pixels upward. Although 2) and 3) are both visually plausible unwarping results, they have large discrepancy in MS-SSIM. However, AD is more reliable and robust to this perturbation. 6) shows dense SIFT flow magnitudes, which are used to compute LD. 85% of the total magnitude of the computed flow is within document margin areas (regions within the red dashed lines), where distortions are imperceptible. Only 2% of the AD flow magnitude in 7) is on the margin.

highly subtle image differences can result in large MS-SSIM discrepancy. LD, on the other hand, is based on dense SIFT flow [Liu et al. 2011], which sometimes generates large flow fields on textureless areas (Fig. 6 b)). These areas are unimportant for document unwarping evaluation but can cause unwanted large error contributions.

To address these issues, we propose Aligned Distortion (AD): a new metric for document unwarping that overcomes the drawbacks of both MS-SSIM and LD. Formally, AD is defined as:

$$AD = \frac{1}{N}\sum_{i=1}^{N} w_i \left\| \mathbf{p_i} - \mathbf{T}^*(\mathbf{p_i} + \mathbf{v_i}) \right\|_2^2, \tag{11}$$

where N is the total number of pixels. $w_i \in [0, 1]$ is the normalized gradient magnitude of pixel i on the document scan. $\mathbf{p_i}$ is the coordinates of the pixel. $\mathbf{v_i}$ is the dense SIFT flow on pixel i. Thus $\mathbf{p_i} + \mathbf{v_i}$ denotes the corresponding pixel coordinate on the unwarped output. $\mathbf{T}^*$ is an optimal constrained affine transformation $\mathbf{T} = \begin{bmatrix} S_x & 0 & T_x \\ 0 & S_y & T_y \\ 0 & 0 & 1 \end{bmatrix}$ such that:

$$\min_{T} \sum_{i} ||\mathbf{p_i} - \mathbf{T}(\mathbf{p_i} + \mathbf{v_i})||_2^2, \forall i \text{ if } w_i > 0.5, \tag{12}$$

where $S_{x,y}$, $T_{x,y}$ are scale and translation respectively. Eq. 12 has a closed form least squares solution.

AD is advantageous: AD is more robust. It aligns the unwarped image and ground truth scan by unifying translation and scale ($\mathbf{T}^*$) before computing the distortion. The alignment step overcomes the drawback of MS-SSIM being sensitive to very subtle global transformations, which generally do not affect the document readability or downstream applications (Fig. 6). AD is more accurate. It weighs the error based on gradient magnitude (w). Compared to $LD = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{v_i}\|_2^2$, this overcomes the drawback of LD which hallucinates errors in textureless regions (Fig. 6). Due to space constraints, we provide a more detailed analysis of these metrics using ablation studies in the supplementary material. However, we provide comprehensive evaluations with experimental results for all metrics, including MS-SSIM, LD, and AD.

We also use OCR performance to evaluate document unwarping quality. We select 45 text-rich images from the benchmark and process them with the Tesseract 4.1.1 OCR engine [Smith 2007] with LSTM backbone. We use the recognition results on the document scans as ground truth, and evaluate OCR performance with Character Error Rate (CER), Word Error Rate (WER), and Edit Distance (ED) [Navarro 2001].

5.2 Baseline Approaches

We compare with 3 model-driven methods: 1) [Tian and Narasimhan 2011], 2) [Kim et al. 2015], and 3) [Kil et al. 2017]. We also compare

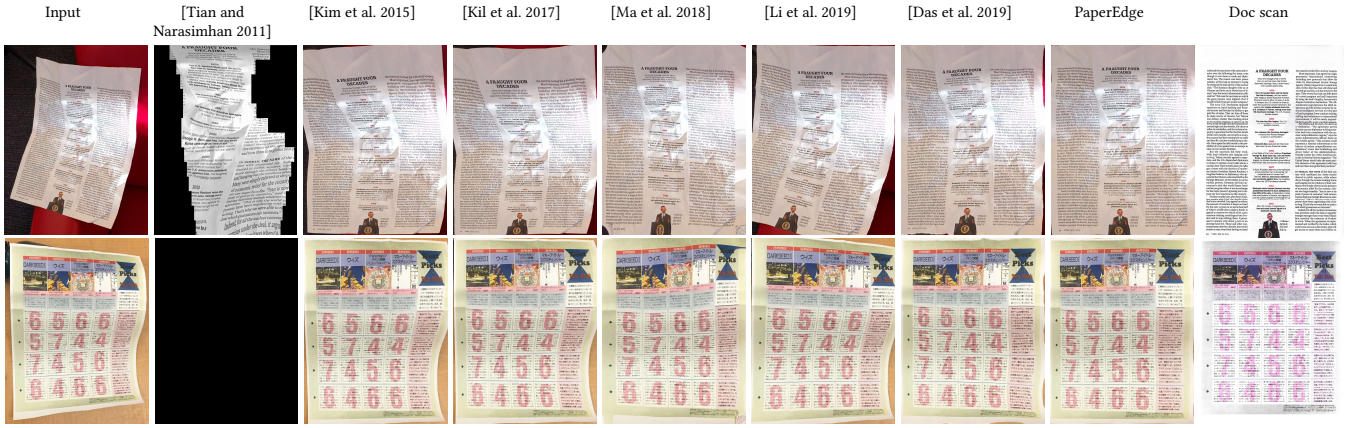

Figure 7: Qualitative unwarping comparison on the benchmark dataset. Our approach generates visually pleasing results with better structure in the content layout and less distorted textlines. Zoom in for details.

with 3 data-driven methods: 4) DocUNet [Ma et al. 2018], 5) [Li et al. 2019], and 6) DewarpNet [Das et al. 2019], which is state-of-art unwarping model. We used code provided by the authors to produce the results. The results from 2), 3), and 5) are not cropped. For a fair comparison, we developed a UNet-based [Ronneberger et al. 2015] auto-cropping system that preserves the largest IoU document area compared to the ground truth document scans for these methods. AD is more robust to inaccurate segmentation because AD aligns the unwarped results to the document scans before computing the distortion. We exclude [Markovitz et al. 2020] from the comparison because neither the code nor the data is available.

5.3 Experimental Results

Learning with only synthetic data. We first demonstrate the effectiveness of our network architecture by training with **only** synthetic data. We train PaperEdge on Doc3D [Das et al. 2019] with ground truth deformation: Enet is trained with L_{SE} in Eq. 2 and Tnet is trained with L_{ST} in Eq. 6. The result is in table 2 as PaperEdge$_D$. Comparing to previous state-of-the-art [Das et al. 2019], PaperEdge achieves slightly better AD and LD, marginally worse but comparable MS-SSIM performance. In OCR evaluation, PaperEdge performs better than [Das et al. 2019] in all related metrics (ED, CER, and WER). This indicates that our network better maintains the readability of these documents after the unwarping.

Learning with both synthetic and real data. The significance of our network design is the capability to utilize both synthetic and real data. To demonstrate the full potential of our method, we train PaperEdge$_{DD}$ with images from both Doc3D and DIW: Enet is trained with $L_{SG} + \lambda L_M$ and Tnet is trained with $L_{SL} + \lambda L_A, \lambda = 0.1$.

In table 2, PaperEdge$_{DD}$ achieves the best results in all image evaluation metrics. PaperEdge$_{DD}$ also achieves the best OCR performance. This demonstrates that our networks are capable of learning from real-world training images and achieves superior generalization ability compared to previous art.

We also provide additional experimental analyses in the supplementary material: We show a categorical analysis on the benchmark dataset to provide better insights on how training with real-world images is especially beneficial for hard-to-synthesis document deformation types. We also present the the long tail pattern of synthetic data and effectiveness of real data.

It is worth noting that, by utilizing a strong Transformer based backbone, [Feng et al. 2021] achieved a comparable AD of 0.398. As to the model size, [Feng et al. 2021] has 26.9M parameters, [Das et al. 2019] is 86.9M, while PaperEdge has 36.4M. We would like to stress that our contributions are orthogonal to [Markovitz et al. 2020] and [Feng et al. 2021]: they are still trained on the synthetic data as our baseline [Das et al. 2019] did. We propose a novel approach to incorporate the real world images in training, which is not explored in the previous work. We expect better performance using stronger backbones such as the architectures from NAS [Zoph and Le 2017] or Transformers.

5.4 Ablation Studies

We perform ablation studies to support our choices on system design and loss functions. Specifically, we analyze individual contributions of each component (Enet and Tnet trained on synthetic and real-world images) using the benchmark dataset.

We evaluate the contribution of each training component via unwarping performance on the DocUNet benchmark [Ma et al. 2018]. To expedite training, we use a slim version of PaperEdge by reducing the intermediate layer depth. The slim PaperEdge has only 14.8M parameters. The result is summarized in table 3. Unwarping with only Enet trained on only synthetic data yields AD of 0.501. With additional 5,000 real images from DIW for training, the performance of Enet increases to 0.481 AD. This improvement demonstrates the effectiveness of training Enet using real-world images. Adding Tnet trained only on synthetic data significantly improves the performance, decreasing AD to 0.424. This improvement is because Tnet fixes local distortions that Enet does not sufficiently address. Additionally, with 5,000 real-world images in training Tnet, our method (PaperEdge$_{DD}$) performs the best among all the training schemes.

Table 3: Ablation studies on different network components. We start evaluating a single Enet trained on Doc3D, then we fine-tune Enet with DIW. After that, we add Tnet trained on Doc3D to the pipeline, and at last, we fine-tune Tnet with DIW. The last configuration is the same as PaperEdge$_{DD}$ in table 2 with a reduced capacity.

Method	AD ↓
Enet (Doc3D)	0.501
Enet (Doc3D + DIW)	0.481
Enet (Doc3D + DIW) + Tnet (Doc3D)	0.424
Enet (Doc3D + DIW) + Tnet (Doc3D + DIW)	**0.413**

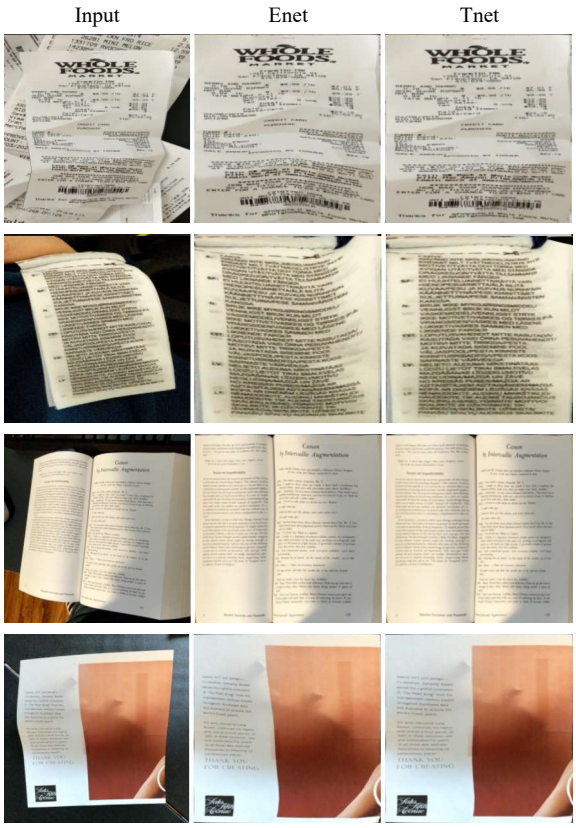

Figure 8: PaperEdge on real world images. In each image triplet, the first image is the input, the second one is the output from Enet, and the last image is the final output of PaperEdge, which is from Tnet. Tnet corrects the local distortion and significantly improves the visual quality.

5.5 Qualitative Results

In Fig. 7, we present the unwarped results on the benchmark and compare with previous methods. [Tian and Narasimhan 2011] fails when text line tracking misses a region, outputting a black area. [Kil et al. 2017] is systematically better than [Kim et al. 2015] as it utilizes more visual cues. Working with patches, [Li et al. 2019] cannot correct large deformations. [Ma et al. 2018] often introduces more distortions in the unwarped images. Our method improves upon DewarpNet [Das et al. 2019] in better reconstructing horizontally straight text lines and vertically straight text columns.

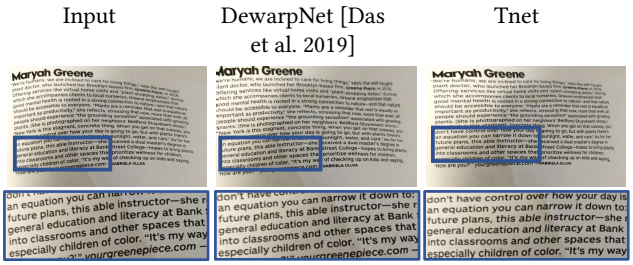

Figure 9: Unwarping partial documents. We compare the PaperEdge Tnet to DewarpNet in unwarping partially visible documents. Tnet outputs more visually appealing results with horizontally straight textlines.

We show more qualitative results from PaperEdge on real-world images in Fig. 8 and compare the intermediate output from Enet to the final refined output from Tnet. The output of Enet presents an edge-based unwarping: the document fills the image plane, and the background is eliminated. Tnet further flattens the remaining distortion in the Enet output. PaperEdge achieves desirable unwarping results on multiple document types in multiple deformations.

In Fig. 9, we demonstrate the utility of the Tnet in unwarping partial documents. Due to texture-based training of Tnet, it can generalize to incomplete document images and better rectify partial documents than previous work.

6 CONCLUSION AND FUTURE WORK

We presented PaperEdge, a novel learning framework for document unwarping. It is the first network that can be trained with both synthetic document images and camera-captured, casual, real world document images. We also proposed Aligned Distortion – a new unwarping performance metric, which overcomes the drawback of previously used MS-SSIM and LD. We collected the Document In the Wild (DIW) dataset of document images in the wild with edge annotations. Training PaperEdge with Doc3D dataset and DIW dataset, we achieve the state-of-the art results in all image similarity metrics and OCR evaluation. In addition, we show that the flexibility of our framework allows it to adapt to new types of data such as partially occluded document images. PaperEdge also has certain limitations because it is a 3D agnostic method. The unwarping result is not guaranteed to be physically correct. PaperEdge also failed to handle some complex crumpled paper cases. With the increasing availability of 3D capturing devices in smartphones in the future, we plan to incorporate a mechanism for learning explicit 3D into PaperEdge to further improve its performance.

ACKNOWLEDGMENTS

This work was done when Ke Ma was at Stony Brook University. This work was partially supported by the Partner University Fund, the SUNY2020 ITSC, and a gift from Adobe.

REFERENCES

John Ashburner. 2007. A Fast Diffeomorphic Image Registration Algorithm. 38, 1 (2007), 95–113.

Adrien Bartoli and Toby Collins. 2013. Template-Based Isometric Deformable 3D Reconstruction with Sampling-Based Focal Length Self-Calibration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Baoquan Chen, Frank Dachille, and Arie Kaufman. 1999. Forward Image Mapping. IEEE.

Ajad Chhatkuli, Daniel Pizarro, and Adrien Bartoli. 2014. Stable Template-Based Isometric 3D Reconstruction in All Imaging Conditions by Linear Least-Squares. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Frédéric Courteille, Alain Crouzil, Jean-Denis Durou, and Pierre Gurdjos. 2007. Shape from Shading for the Digitization of Curved Documents. *Machine Vision and Applications* 18, 5 (2007), 301–316.

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 International Conference on Computer Vision*.

Sagnik Das, Gaurav Mishra, Akshay Sudharshana, and Roy Shilkrot. 2017. The Common Fold: Utilizing the Four-Fold to Dewarp Printed Documents from a Single Image. In *Proceedings of the 2017 ACM Symposium on Document Engineering (DocEng '17)*. 125–128. https://doi.org/10.1145/3103010.3121030

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 International Conference on Computer Vision*.

Hironori Ezaki, Seiichi Uchida, Akira Asano, and Hiroaki Sakoe. 2005. Dewarping of Document Image by Global Optimization. In *Proceedings of the International Conference on Document Analysis and Recognition*.

Hao Feng, Yuechen Wang, Wengang Zhou, Jiajun Deng, and Houqiang Li. 2021. DocTr: Document Image Transformer for Geometric Unwarping and Illumination Correction. In *Proceedings of the ACM International Conference on Multimedia*.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. 2018. Unsupervised Representation Learning by Predicting Image Rotations. In *Proceedings of International Conference on Learning and Representation*.

Nail Gumerov, Ali Zandifar, Ramani Duraiswami, and Larry S Davis. 2004. Structure of Applicable Surfaces from Single Views. In *Proceedings of the European Conference on Computer Vision*. Springer, 482–496.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

David Fuentes Jiménez, David Casillas Pérez, Daniel Pizarro Pérez, Toby Collins, and Adrien Bartoli. 2018. Deep Shape-from-Template: Wide-Baseline, Dense and Fast Registration and Deformable Reconstruction from a Single Image. *arXiv preprint arXiv:1811.07791* (2018). arXiv:1811.07791

Rahat Khan, Daniel Pizarro, and Adrien Bartoli. 2014. Schwarps: Locally Projective Image Warps Based on 2d Schwarzian Derivatives. In *Proceedings of the European Conference on Computer Vision*. Springer.

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*. IEEE, 865–870.

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.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese Neural Networks for One-Shot Image Recognition. In *ICML Deep Learning Workshop*. Lille.

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)* (2019).

Jian Liang, Daniel DeMenthon, and David Doermann. 2008. Geometric Rectification of Camera-Captured Document Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 4 (2008), 591–605.

Chen-Hsuan Lin and Simon Lucey. 2017. Inverse Compositional Spatial Transformer Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2568–2576.

Ce Liu, Jenny Yuen, and Antonio Torralba. 2011. Sift Flow: Dense Correspondence across Scenes and Its Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 5 (2011), 978–994.

Changsong Liu, Yu Zhang, Baokang Wang, and Xiaoqing Ding. 2015. Restoring Camera-Captured Distorted Document Images. *International Journal on Document Analysis and Recognition* 18, 2 (2015), 111–124.

Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. 2018. An Intriguing Failing of Convolutional Neural Networks and the Coordconv Solution. *arXiv preprint arXiv:1807.03247* (2018). arXiv:1807.03247

Xiyan 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).

Shijian Lu and Chew Lim Tan. 2006. Document Flattening through Grid Modeling and Regularization. In *Proceedings of the International Conference on Pattern Recognition*.

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*.

Amir Markovitz, Inbal Lavi, Or Perel, Shai Mazor, and Roee Litman. 2020. Can You Read Me Now? Content Aware Rectification Using Angle Supervision. In *Proceedings of the European Conference on Computer Vision*. Springer.

Gaofeng Meng, Zuming Huang, Yonghong Song, Shiming Xiang, and Chunhong Pan. 2015. Extraction of Virtual Baselines from Distorted Document Images Using Curvilinear Projection. In *Proceedings of the International Conference on Computer Vision*.

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*.

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*.

Gonzalo Navarro. 2001. A Guided Tour to Approximate String Matching. *ACM computing surveys (CSUR)* 33, 1 (2001), 31–88.

Albert Pumarola, Antonio Agudo, Lorenzo Porzi, Alberto Sanfeliu, Vincent Lepetit, and Francesc Moreno-Noguer. 2018. Geometry-Aware Network for Non-Rigid Shape Prediction from a Single View. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention*.

Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. 2004. " GrabCut" Interactive Foreground Extraction Using Iterated Graph Cuts. *ACM Transactions on Graphics (TOG)* 23, 3 (2004), 309–314.

Faisal Shafait and Thomas M Breuel. 2007. Document Image Dewarping Contest. In *Workshop on Camera-Based Document Analysis and Recognition*.

R. Smith. 2007. An Overview of the Tesseract OCR Engine. In *Proceedings of the International Conference on Document Analysis and Recognition*. 5 pages.

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*.

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*.

Adrian Ulges, Christoph H. Lampert, and Thomas Breuel. 2004. Document Capture Using Stereo Vision. In *Proceedings of the 2004 ACM Symposium on Document Engineering (DocEng '04)*. 198–200. https://doi.org/10.1145/1030397.1030434

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*.

VSR Veeravasarapu, Abhishek Goel, Deepak Mittal, and Maneesh Singh. 2020. ProAlignNet: Unsupervised Learning for Progressively Aligning Noisy Contours. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9671–9679.

Toshikazu Wada, Hiroyuki Ukida, and Takashi Matsuyama. 1997. Shape from Shading with Interreflections under a Proximal Light Source: Distortion-free Copying of an Unfolded Book. *International Journal of Computer Vision* 24, 2 (1997), 125–135.

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 *Document Analysis Systems*. Springer, 131–144.

Chuhui Xue, Zichen Tian, Fangneng Zhan, Shijian Lu, and Song Bai. 2022. Fourier Document Restoration for Robust Document Dewarping and Recognition. *arXiv preprint arXiv:2203.09910* (2022). arXiv:2203.09910

Shaodi You, Yasuyuki Matsushita, Sudipta Sinha, Yusuke Bou, and Katsushi Ikeuchi. 2017. Multiview Rectification of Folded Documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. 2019. Aet vs. Aed: Unsupervised Representation Learning by Auto-Encoding Transformations Rather than Data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2547–2555.

Barret Zoph and Quoc V. Le. 2017. Neural Architecture Search with Reinforcement Learning. In *Proceedings of International Conference on Learning and Representation*.