Fourier Document Restoration for Robust Document Dewarping and Recognition

Chuhui Xue[1], Zichen Tian[1], Fangneng Zhan[1], Shijian Lu[1], Song Bai[2]
[1]Nanyang Technological University, [2]ByteDance
xuec0003@e.ntu.edu.sg, {zichen.tian,shijian.lu,fnzhan}@ntu.edu.sg, songbai.site@gmail.com

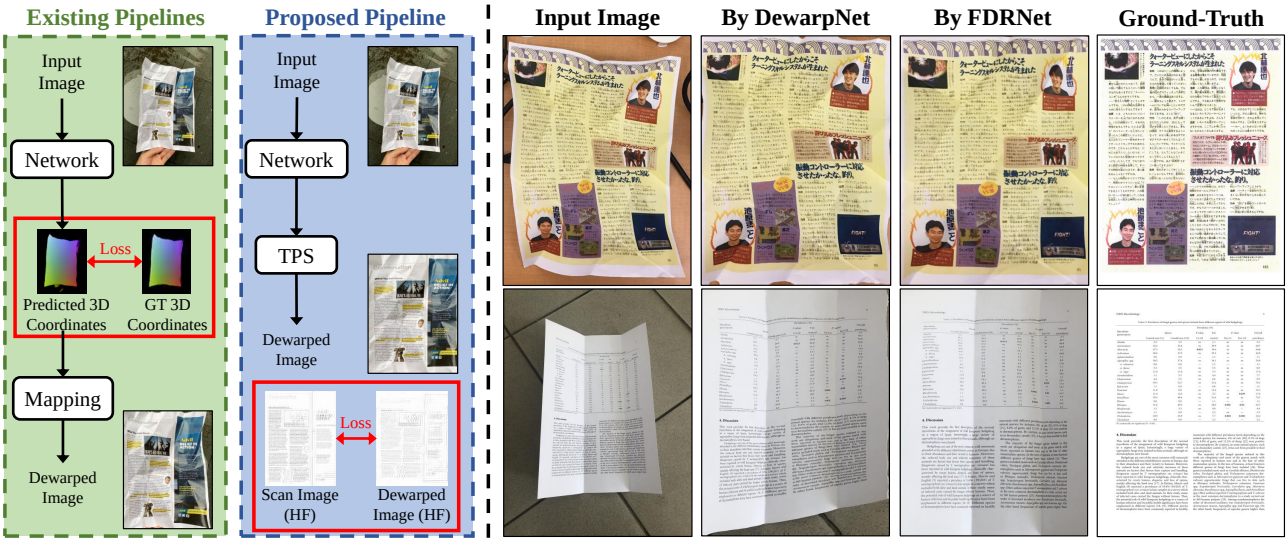

Figure 1. **Existing document dewarping and the proposed FDRNet:** Existing document dewarping learns to predict 3D coordinates of each pixel in camera document for dewarping, which often struggle when handling documents with irregular distortions or large depth variations as shown in column 2 in the right graph. FDRNet instead focuses on the high-frequency components of document contents and learns to dewarp the whole document with Thin-Plate Spline (TPS) transformation. It is robust to irregular deformations and depth variations as shown in column 3 in the right graph, and requires much less simply annotated training data.

Abstract

State-of-the-art document dewarping techniques learn to predict 3-dimensional information of documents which are prone to errors while dealing with documents with irregular distortions or large variations in depth. This paper presents FDRNet, a Fourier Document Restoration Network that can restore documents with different distortions and improve document recognition in a reliable and simpler manner. FDRNet focuses on high-frequency components in the Fourier space that capture most structural information but are largely free of degradation in appearance. It dewarps documents by a flexible Thin-Plate Spline transformation which can handle various deformations effectively without requiring deformation annotations in training. These features allow FDRNet to learn from a small amount of simply labeled training images, and the learned model can dewarp documents with complex geometric distortion and recognize the restored texts accurately. To facilitate document restoration research, we create a benchmark dataset consisting of over one thousand camera documents with different types of geometric and photometric distortion. Extensive experiments show that FDRNet outperforms the state-of-the-art by large margins on both dewarping and text recognition tasks. In addition, FDRNet requires a small amount of simply labeled training data and is easy to deploy. The proposed dataset is available at https://sg-vilab.github.io/event/warpdoc/.

1. Introduction

Automated document recognition is critical in many applications such as library digitization, office automation, e-

business, etc. It has been well solved by optical character recognition (OCR) technology if documents are properly scanned by document scanners. But for increasing document images captured by various camera sensors, OCR software often encounters various recognition problems due to two major factors. First, document texts captured by cameras often lie over a curved or folded surface and suffer from different types of geometric distortions such as document warping, folding, and perspective views as illustrated in Fig. 1. Second, document texts captured by cameras often suffer from different types of photometric distortion due to uneven illuminations, motion, shadows, etc. Accurate recognition of document texts captured by camera sensors remains a grand challenge in the document analysis and recognition research community.

Document restoration has been investigated extensively for better recognition of documents captured by various camera sensors. Recent data-driven methods [8, 11] synthesize 3D document images with various distortions and learn document distortions by predicting the 3D coordinates of *each pixel* in warped documents which have achieved very impressive performances on document dewarping task. However, these methods are facing three challenges. First, most pixels in document images suffer from regular distortions of perspective or curvature, whereas only a small portion of pixels exhibit irregular deformations (e.g. pixels around crumples). Such pixel-level data imbalance often leads to degraded performance for existing pixel-level regression-based models while handling documents with irregular deformations as shown in the first row on the right of Fig. 1. Second, most existing document dewarping methods perform poorly when documents are far away from the camera as illustrated in the second row on the right of Fig. 1. This is largely because existing methods often struggle in predicting document 3D coordinates when the document depth has large variations. Third, most existing models are trained on large amounts of synthetic images, where the synthesis is complicated requiring to collect 3D coordinates by special hardware (i.e. depth camera) and a large number of scanned document images (i.e. 100,000 synthetic images from 3D coordinates of 1,000 documents and 7,200 scanned images in [8]). This makes it challenging to generalize existing methods to new tasks and domains.

We design FDRNet, an end-to-end trainable document restoration network that focuses on document contents and aims for better document recognition. FDRNet is inspired by the observation that geometric distortions in document images can be largely inferred from high-frequency components in Fourier space whereas appearance degradation is largely encoded in low-frequency components. Document restoration and recognition should therefore focus on high-frequency components capturing document structures and contents and ignoring interfering low-frequency components capturing largely appearance noises. We thus design FDRNet to learn geometric distortions by focusing on high-frequency information of the whole document instead of 3D coordinates of each pixel which helps tackle the challenge of pixel-level data imbalance and document depth variation effectively. FDRNet is powered by Thin-Plate Spline transformation which helps not only reduce training data significantly but also eliminate the need for 3D document coordinates ground-truthing and the complex data collection process in training. Furthermore, we introduce WarpDoc, a benchmarking dataset with more than one thousand document images with different types of degradation in geometry and appearance that greatly help for better validation of document dewarping models. Extensive experiments show that FDRNet achieves superior document restoration as illustrated in Fig. 1.

The contributions of this work are three-fold. First, we design FDRNet, an end-to-end trainable document restoration network that can remove geometric and appearance degradation from camera images of documents and improve document recognition significantly. Second, FDRNet handles document restoration and recognition by focusing on high-frequency components in the Fourier space which helps reduce training data and improve model generalization and usability greatly. Third, we create a dataset with more than one thousand camera images of documents which is very valuable to future research in the restoration and recognition of documents captured by cameras.

2. Related Works

2.1. Geometric Document Restoration

Document texts captured by camera sensors often lie over a curved/folded surface and suffer from various perspective distortions that hinder document recognition significantly. Document dewarping has been studied extensively for flattening documents into a recognition-friendly form. Traditional methods dewarp documents by reconstructing 3D document shapes [5, 15, 29–31, 37, 39, 44, 45, 48] or extracting 2D image features [6, 10, 13, 17, 20, 22, 24, 25, 28, 34, 38, 40, 41, 47]. On the other hand, extracting 2D features often involves various heuristic parameters and 3D reconstruction is complicated and sensitive to various noises. In recent years, some work [8, 9, 27] exploits deep neural networks to learn document shapes from 2D/3D synthetic document images. However, such data-driven methods require a large amount of synthetic data that are complicated and time-consuming to collect.

Our proposed FDRNet dewarps documents by learning 2D deep network features with little heuristics. Instead of using large amounts of synthetic data [8, 9, 11, 27], it learns from high-frequency components of real document images which allows learning superior geometric document

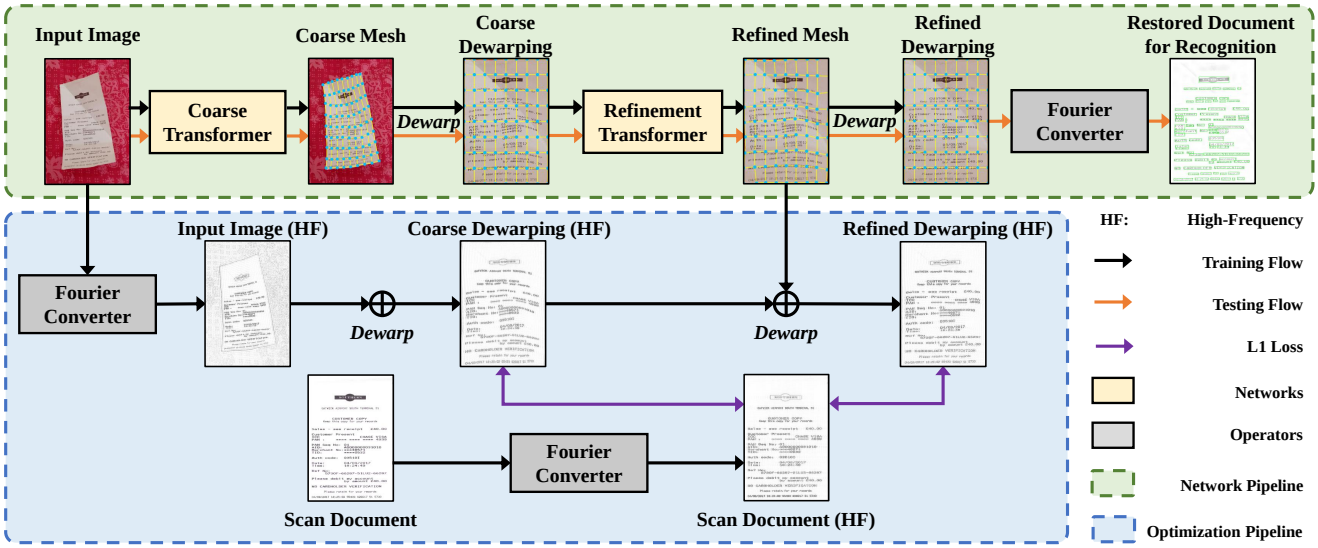

Figure 2. **The framework of the proposed FDRNet**: With camera captured *Input Image*, FDRNet learns to predict control points (for thin-plate-spline transformation in dewarping) by a *Coarse Transformer* and a *Refinement Transformer* and use the predicted control points as the node of *Coarse Mesh* and *Refined Mesh* for document dewarping. It computes rectification losses (L_1 loss) over high-frequency information of the *Input Images* as produced by a *Fourier Converter* and the corresponding *Scanned Documents* without using any annotations in training. In inference, the dewarped document is fed to a *Fourier Converter* for photometric restoration and recognition.

restoration models with a small amount of training data.

2.2. Photometric Document Restoration

Document images captured by camera devices often suffer from various illumination noises such as occlusion shadows as induced by photographers or documents themselves. Such illumination noises complicate text segmentation from document background, which could degrade text recognition performance significantly. Different photometric restoration and document image binarization techniques [1, 18, 19, 23, 36] have been reported for segmenting texts from various unevenly illuminated document images. On the other hand, most existing works are either computationally intensive [7, 36] or sensitive to heuristic parameters [1, 3, 23, 26, 35] and not a good fit as a pre-processing step of document recognition. More recently, some approaches [11, 16] correct illuminations of documents by patch-based networks. Our proposed technique handles illumination noises by extracting high-frequency document information, which is efficient and robust and involves minimal heuristics.

3. Methodology

The proposed FDRNet consists of three components including a *Coarse Transformer*, a *Refinement Transformer*, and a *Fourier Converter* as illustrated in Fig. 2. The *Coarse Transformer* and *Refinement Transformer* learn to dewarp documents in a coarse-to-fine manner. The *Fourier Converter* extracts high-frequency information of document images for effective and efficient network training as shown in the *Optimization Pipeline* as highlighted in green in Fig. 2. Additionally, it extracts high-frequency content information for better document recognition as shown at the right end of the *Network Pipeline* as highlighted in blue in Fig. 2.

3.1. Coarse-To-Fine Transformer

FDRNet dewarps document images in a coarse-to-fine manner by using a Coarse Transformer and a Refinement Transformer. The two transformers share the same architecture Spatial Transformer Network (STN) [14] that models the spatial transformation as learnable networks. Specifically, the Coarse Transformer learns to localize the document region in the input image and dewarps the located document region coarsely. The Refinement Transformer takes the dewarped document image from the Coarse Transformer and improves the dewarping further.

We adopt Thin-Plate-Spline [4] (TPS) as the spatial transformation in document dewarping. TPS transformation is determined by two sets of control points with a one-to-one correspondence between a pair of warped and flat document images, and it computes a spatial deformation function for every control point to predict geometric distortions. In FDRNet, we define the control points as mesh grid and the network learns to predict the mesh grid of document region in the input image (i.e. the blue dots in *Predicted Mesh* in Fig. 2). With the predicted mesh grid, TPS transforms them to the regular mesh grid (i.e. the blue dots

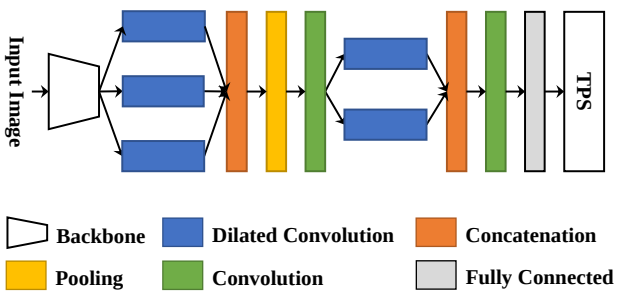

Figure 3. The **Architecture** of *Coarse Transformer* and *Refinement Transformer* for coarse-to-fine document distortion estimation and rectification (TPS: Thin-Plate-Spline)

in *Coarse Dewarping* and *Refined Dewarping* in Fig. 2) to achieve document dewarping. The mesh grid can have different sizes and our study shows that a 9×9 mesh grid (with 81 control points) is sufficient for document dewarping.

By denoting the predicted mesh grid points by $P = [\boldsymbol{t_1}, \boldsymbol{t_2}, ..., \boldsymbol{t_k}]^T$ and the regular mesh grid points by $P' = [\boldsymbol{t'_1}, \boldsymbol{t'_2}, ..., \boldsymbol{t'_k}]^T$, the TPS transformation parameters can be determined as follows:

$$C_x = \begin{bmatrix} S & 1_k & P \\ 1_k^T & 0 & 0 \\ P^T & 0 & 0 \end{bmatrix}^{-1} \cdot \begin{bmatrix} P'_x \\ 0 \\ 0 \end{bmatrix}, \quad (1)$$

where each element $(S)_{ij}$ in S is determined by $\phi(\boldsymbol{t_i} - \boldsymbol{t_j})$ and $\phi(\boldsymbol{r})$ is defined by $\|\boldsymbol{r}\|_2^2 \, log \, \|\boldsymbol{r}\|_2^2$. P'_x refer to x coordinates of P'. Similarly, C_y can be obtained by replacing P'_x by P'_y. Hence, we can get $C = [C_x, C_y]$. Finally, for each control point of the document region in the input image $\boldsymbol{u}$, the corresponding point $\boldsymbol{u'}$ in the dewarped document can be determined by:

$$\boldsymbol{u}' = C \cdot \boldsymbol{u}. \quad (2)$$

Note that the predicted mesh grid points are initialized by regular mesh grid in the implementation. Since all operators in the TPS transformation are differentiable, the *Coarse Transformer* and *Refinement Transformer* can learn to localize document mesh grid points by gradient backpropagation without requiring any annotation of document mesh grids. Additionally, we adopt stacked dilated convolution [32, 46] in the two transformers to enlarge the network receptive field since the mesh grid localization requires to focus on high-level document content information. Fig. 3 shows the detailed structure of the Coarse and Refinement Transformers. Specifically, document features are first extracted by a backbone network which are then fed to three stacked dilated convolutional layers followed by two stacked dilated convolutional layers of different dilation rates [42]. The network finally predicts a set of control points (as the document mesh grid as illustrated in *Predicted Mesh* in Fig. 2) and passes them to TPS for document dewarping.

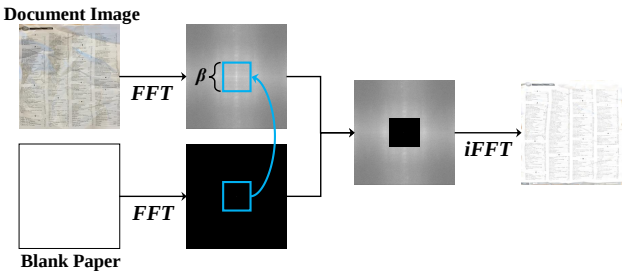

Figure 4. **Illustration of the proposed Fourier Converter**: The Fourier Converter transforms a document image and a blank paper image into Fourier space by FFT. The low-frequency components of the document image are then replaced by the corresponding components of the blank paper (as highlighted by blue boxes). The modified spectral signals are finally transformed back to spatial space by iFFT, where most low-frequency appearance noises are removed with little effects over high-level content information.

3.2. Fourier Converter

We design a Fourier Converter to extract high-frequency information from document images captured by cameras. Given a *Document Image* as shown in Fig. 4, the Fourier Converter first transforms it into Fourier space via Fast Fourier Transform (FFT) [12]. Next, the document's low-frequency information is replaced with the low-frequency information of a *Blank Paper*. The modified spectral signals are finally transformed back to the spatial space (through inverse Fast Fourier Transform (iFFT)), which produces the OCR-friendly document images with most appearance noises successfully removed.

We employ a hyper-parameter β in Fourier Converters in both network training and document recognition tasks. As Fig. 4 shows, β controls how much low-frequency information (the center has the lowest frequency) is replaced. The high-frequency information can thus be extracted with a mask M_β of size (H, W) as follows:

$$M_\beta(h, w) = \begin{cases} 0, & (h, w) \in [-\beta H : \beta H, -\beta W : \beta W] \\ 1, & Otherwise \end{cases},$$

where $\beta \in [0, 1/2]$, $h \in [-H/2, H/2]$ and $w \in [-W/2, W/2]$. With x denoting the spectral signals of the dewarped document and x_w denoting that of the blank paper, the modified spectral signals can be derived as follows:

$$x' = M_\beta \cdot x + (1 - M_\beta) \cdot x_w. \quad (3)$$

We empirically set β at 0.06 and 0.008, respectively, for the two Fourier Converters in network training and document recognition. Since β is a ratio in Fourier space that is invariant to image sizes or resolutions, it can be directly applied to various new images with little tuning, more details to be discussed in the ensuing Experiments.

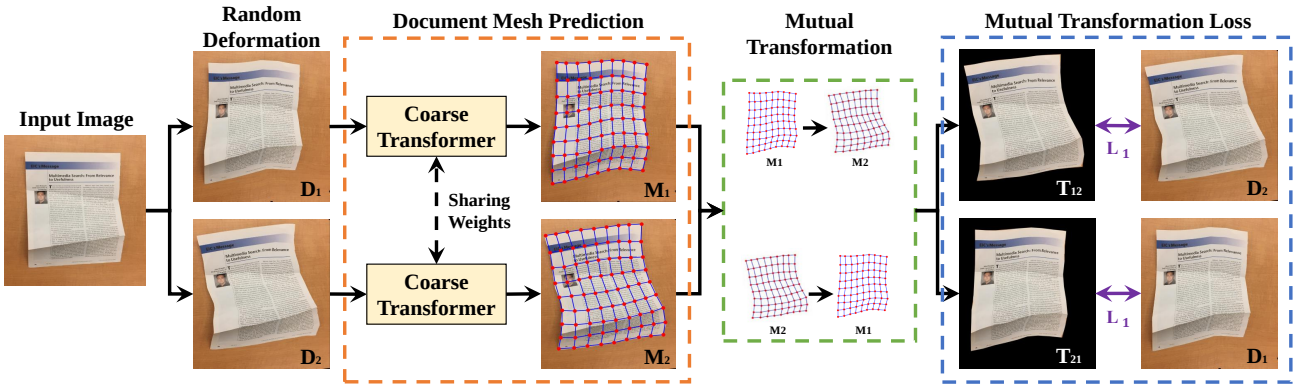

Figure 5. **Illustration of our proposed Mutual Transformation Loss**: Each *Input Image* is randomly deformed to two distorted images D_1 and D_2. FDRNet then learns to predict mesh grids of the document region M_1 and M_2 within the two distorted images, and transform them to each other mutually which produces T_{12} and T_{21}. The difference between each distorted image and its transformation is computed to guide FDRNet to better focus on document distortion in training.

The Fourier Converter helps to train FDRNet effectively. Given the *Input Image* and the corresponding *Scanned Document*, it first extracts high-frequency information *Input Image (HF)* and *Scanned Document (HF)* as illustrated in Fig. 2. At each training batch, the *Input Image (HF)* is dewarped to produce *Coarse Dewarping (HF)* and *Refined Dewarping (HF)* by the *Coarse Transformer* and the *Refinement Transformer*, respectively. FDRNet learns by back-propagating $\mathcal{L}_1$ loss between the *Scan Document (HF)* and the *Coarse Dewarping (HF)* & *Refined Dewarping (HF)*. The Fourier Converter improves network learning from two aspects. First, it discards low-frequency appearance information that often contains rich noisy variations and makes network learning much more complicated. Thanks to such data cleaning, FDRNet can be trained effectively and efficiently by using a small amount of training data. Second, the clear appearance gap between camera-captured documents and scanned documents often affects the stability and convergence of network training. Fourier Converter extracts high-frequency information which minimizes the domain gaps and enables direct loss computation between the two types of document images without requiring any annotations of mesh grids in training.

For document recognition during the inference stage, the Fourier Converter extracts high-frequency information from the *Refined Dewarping* which often suffers from various appearance noises that degrade the document recognition performance greatly. This removes various appearance noises effectively and improves the document recognition greatly as illustrated in Fig. 2.

3.3. Network Training

FDRNet can be trained by optimizing the *Coarse Transformer* and the *Refinement Transformer* only as the *Fourier Converter* is frozen with empirically determined β in training. We train the *Coarse Transformer* by using a L_{rect} loss and a mutual transformation loss as follows:

$$\mathcal{L}_{coarse} = \mathcal{L}_{rect} + \lambda * \mathcal{L}_{mutual}, \tag{4}$$

where $\mathcal{L}_{rect}$ is L_1 loss that can be directly computed between the *Coarse Dewarping (HF)* and *Scan Document (HF)* as illustrated in Fig. 2. The L_1 loss works well as *Coarse Dewarping (HF)* and *Scan Document (HF)* have similar intensity but little appearance noises and domain gaps. Parameter λ is the weight to balance the two losses which is empirically set at 0.5 in our network.

Since document images captured by cameras often suffer from severe geometric distortions, the network training may not converge (with L_1 loss alone) without ground-truth annotations of document meshes. We design a mutual transformation loss L_{mutual} that 'fabricates' certain supervision to constrain and guide the network to learn geometric distortions stably. The underlying idea of L_{mutual} is that a document with two different geometric distortions can be mutually transformed to each other if their mesh grids are predicted correctly. In implementation, the *Input Image* is first transformed to two new images (i.e. D_1 and D_2) with randomly perturbed deformation [27] as illustrated in Fig. 5. The two transformed images are then fed to FDRNet to predict the corresponding document meshes M_1 and M_2, respectively. D_1 can thus be transformed to image T_{12} by TPS transformation $M_1 \rightarrow M_2$, and D_2 can be similarly transformed to image T_{21} by $M_2 \rightarrow M_1$. The mutual transformation loss is thus defined as follows:

$$\mathcal{L}_{mutual} = \|T_{12} - D_2\| * m_2 + \|T_{21} - D_1\| * m_1, \tag{5}$$

where m_1 and m_2 refer to document regions within M_1 and M_2. Note although perturbed distortion could produce abnormal distortions around the document background, FDR-

Deformation	Description
Perspective	With perspective distortion only
Fold	With one or several creases on document
Curved	With curvature distortion
Random	With random crumples
Rotating	With in-plane rotation between -45 and 45
Incomplete	With incomplete distortion without affecting the content of documents.

Table 1. **Details of the proposed WarpDoc Benchmark:** The WarpDoc Benchmark contains 1,020 camera captured document images with six different types of deformation including Perspective, Fold, Curved, Random, Rotating and Incomplete.

Methods	Training Data			MS-SSIM	LD
	No.	D-GT	Type		
DocUNet [27]	100k	✓	Synth	0.41	14.08
GBSUM [2]	8k	✓	Synth	0.42	13.20
AGUN [21]	40k	✓	Synth	0.45	12.06
DewarpNet [8]	100k	✓	Synth	0.47	8.95
DocTr [11]	100k +3k	✓	Synth +Real	**0.50**	**8.38**
FDRNet	1k	✗	Real	**0.50**	9.43

Table 2. **Image similarity (in MS-SSIM and LD) over DocUNet:** No.: Number of training images; D-GT: Deformation Ground-Truth; Synth: Synthetic images; Real: Real images.

Net can focus on document regions progressively by restricting loss computation within document meshes and ignoring the document background simultaneously.

The *Refinement Transformer* can be trained by using an L_1 loss alone (in between *Refined Dewarping (HF)* and *Scan Document (HF)* as shown in Fig. 2). It does not require the mutual transportation loss as the *Coarse Transformer* has located document regions and rectified most geometric distortions. The L_1 loss alone is sufficient for the prediction of the remaining document distortion.

4. Experiments

4.1. Datasets

We evaluated FDRNet over two datasets as listed:

DocUNet [27]: DocUNet contains 130 images that are taken for different paper documents with different contents and texts of different languages. These images are taken under different conditions which suffer from various distortions. For each paper document, a scanned copy is collected as a ground-truth document.

WarpDoc: We collected WarpDoc, a warped document image dataset for evaluating document restoration methods. WarpDoc consists of 1,020 camera images of documents that were collected from scientific papers, magazines, envelopes, etc., which have different paper materials, page layouts, and contents. The images were taken in different scenes (indoors, outdoors, etc.) with different illuminations. Before imaging, we warped the 1,020 printed documents into six types of distortions including Fold, Curved, Random, Rotating, Incomplete, and Perspective as illustrated in columns 3-8 of Fig. 6, respectively. More details about our WarpDoc are available in the Supplementary Material.

Methods	Training Data			MS-SSIM	LD
	No.	D-GT	Type		
GBSUM [2]	8k	✓	Synth	0.34	29.07
DewarpNet [8]	100k	✓	Synth	0.33	31.15
FDRNet	130	✗	Real	**0.45**	**20.30**
GBSUM-Crop [2]	8k	✓	Synth	0.41	23.34
DewarpNet-Crop [8]	100k	✓	Synth	0.39	21.89
FDRNet-Crop	130	✗	Real	**0.46**	**19.11**

Table 3. **Image similarity (in MS-SSIM and LD) over WarpDoc:** Crop: Evaluation on tightly cropped images from WarpDoc Benchmark; No.: Number of training images; D-GT: Deformation Ground-Truth; Synth: Synthetic images; Real: Real images.

Methods	CER(%)	
	DocUNet Benchmark	WarpDoc Benchmark
GBSUM [2]	37.94	66.48
DewarpNet [8]	23.95	45.82
DocTr [11]	20.00	-
FDRNet	**16.96**	**29.24**

Table 4. **Character error rates** over DocUNet and WarpDoc.

4.2. Evaluation Metrics

We adopt two types of widely used evaluation metrics [8, 11, 27, 45] including: 1) Multi-Scale Structural Similarity (MS-SSIM) [43] and Local Distortion (LD) [45] that focus on image similarity performance; 2) Character Error Rate (CER) for evaluation of optical character recognition (OCR) performance. More details are available in the Supplementary Material.

4.3. Experimental Results

We conduct cross-validation experiments over DocUNet and WarpDoc benchmarks for evaluation of FDRNet qualitatively and quantitatively. For each test document image, FDRNet model produces two images including a dewarped document image with geometric restoration only and a fully restored document image with further appearance restoration as illustrated in rows 3 and 4 in Fig. 6, respectively. We evaluate the dewarped document images by using image similarity metrics and the appearance-restored document image by using OCR accuracy.

Image Similarity: Tab. 2 shows the MS-SSIM and LD of the proposed FDRNet as well as several state-of-the-art methods over DocUNet and WarpDoc. As Tab. 2 shows,

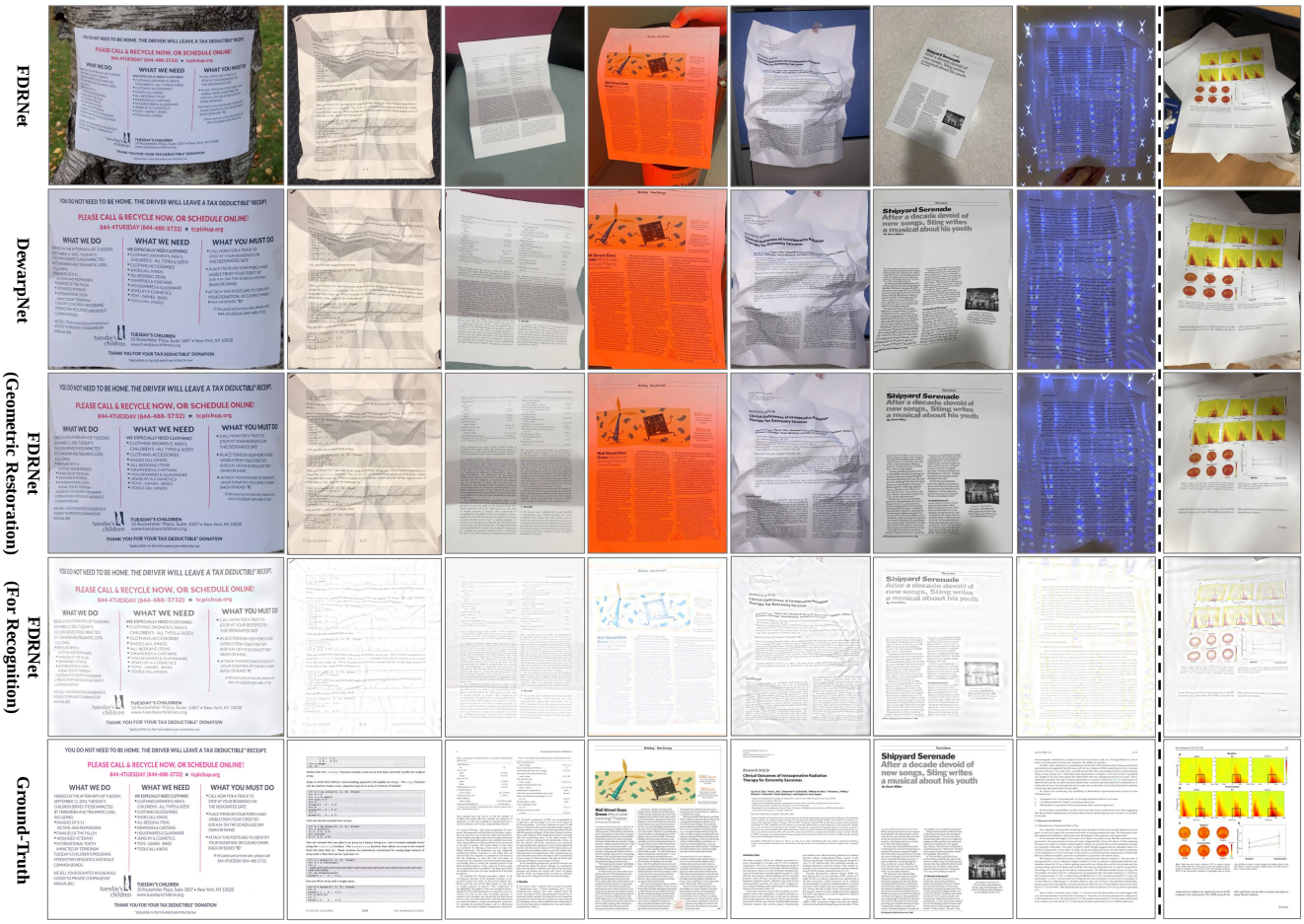

Figure 6. **Illustration of document restoration by FDRNet and DewarpNet**: For the sample images from DocUNet in columns 1-2 and WarpDoc in columns 3-8 in the first row, rows 2 and 3 show the dewarped images by using DewarpNet and FDRNet (dewarping on), respectively. Row 4 shows the appearance restoration by FDRNet which removes various appearance noises and improves document recognition greatly. FDRNet is robust to most geometric and photometric distortions but tends to get confused while document background has similar patterns as document regions as illustrated in the last sample.

FDRNet achieves competitive dewarping performance over the DocUNet. On the other hand, FDRNet uses much simpler training data than state-of-the-art methods in both image number (1k v.s. 8k-100k) and image annotations (w/o v.s. w/ deformation ground-truth).

We further evaluate FDRNet on the proposed WarpDoc dataset in which document images usually suffer from much more complex distortions than document images in DocUNet benchmark. We conduct two sets of experiments for a better comparison with the state-of-the-art. First, we compare FDRNet with existing document dewarping methods on the original WarpDoc dataset to evaluate document dewarping under the presence of both complex geometric distortions and significant depth variation. Second, we crop the images in the WarpDoc dataset (following [27]) to reduce the depth variation of documents in the original images. We hence compare FDRNet with existing methods on the cropped images in which depth variations of documents are largely mitigated. As Tab. 3 shows, the proposed FDRNet outperforms the existing approaches on dewarping documents with complex geometric distortions alone or additional depth variation by using much fewer and simpler training samples. This result shows that the proposed FDRNet is more robust to document dewarping as compared with the state-of-the-art. Additionally, the GBSUM and DewarpNet achieve very different performances on dewarping original and cropped document images, showing that they are sensitive to document depth variations. On the contrary, the dewarping performance of FDRNet on original and cropped images is similar, demonstrating the proposed FDRNet is much more robust to the depth variation of documents as compared with existing approaches.

FDRNet Components					Experimental Results		
CT	FC_{tr}	*MTL*	*RT*	FC_{inf}	MS-SSIM	LD	CER(%)
✓					Not converge		
✓	✓				0.32	34.16	69.32
✓	✓		✓		0.37	23.47	48.24
✓	✓	✓			0.44	16.35	33.02
✓	✓	✓	✓		**0.50**	**9.43**	23.46
✓	✓	✓	✓	✓	-	-	**16.96**

Table 5. **Ablation study** of FDRNet over DocUNet: FC_{tr} - Fourier Converter for training; *CT* - Coarse Transformer; *MTL* - Mutual Transformation Loss; *RT* - Refinement Transformer; FC_{inf} - Fourier Converter for inference.

Fig. 6 shows the restoration of several sample images from DocUNet and WarpDoc that suffer from different types of distortions. As Fig. 6 shows, FDRNet achieves similar restoration as DewarpNet for documents with simple curvature distortions (sample in column 1). But for documents with more complex distortions in columns 2-7, FDRNet usually performs better as it focuses on high-frequency information where document layouts such as text lines help to learn geometric distortions better. As a comparison, 3D methods such as DewarpNet regress each pixel from warped documents to flat ones. The regression of pixels around complex crumples is often hard to learn as there are much fewer such pixels as compared with those with simple distortions. FDRNet learns a general transformation by a coarse mesh grid which is less affected by the pixel-level data imbalance during network training.

OCR Performances: We examine how FDRNet performs on document recognition by evaluating OCR over FDRNet restored documents using PyTesseract (v4.1.1) [33]. Following DewarpNet [8], we perform OCR over 54 document images on DocUNet and 739 document images on WarpDoc with lots of texts. Tab. 4 shows experimental results. We can observe that FDRNet achieves CER of 16.96% and 29.24% on DocUNet and WarpDoc, respectively, which outperforms state-of-the-art methods with illumination restoration by large margins. More specifically, although the performances of FDRNet and state-of-the-art methods are comparable on the metrics of image similarity on DocUNet dataset as shown in Tab. 2, FDRNet outperforms these approaches by a large margin on CER, demonstrating that FDRNet is more robust to document recognition task. The second last row in Fig. 6 illustrates the FDRNet restored documents images. It can be seen that FDRNet removes various geometric and appearance distortions from the dewarped documents which facilitate OCR and document recognition significantly.

4.4. Discussion

Ablation studies: We study the contributions of different designs in our FDRNet including a Fourier Converter for network training FC_{tr}, a Coarse Transformer *CT*, a Mutual Transformation Loss *MTL*, a Refinement Transformer *RT* and a Fourier Converter for inference FC_{inf}. Tab. 5 shows the experimental results.

As shown in Tab. 5 rows 1-2, the *CT* alone cannot converge due to unstable losses during training that are caused by the large domain gap between document images collected by cameras and scanners. By including the proposed FC_{tr}, FDRNet training stabilizes. The further inclusion of *MTL* and *RT* both help to train more powerful dewarping models with clearly improved MS-SSIM and LD, as shown in rows 3-5. During inference, including the Fourier Converter (i.e. FC_{inf}) improves OCR by large margins as FC_{inf} removes various appearance noises that often affect document recognition, as shown in row 6.

β	0.003	0.005	0.008	0.01	0.02
CER(%)	18.52	17.84	16.96	17.38	17.72

Table 6. CER varies with the parameter β in the Fourier Converter (described in Section 3.2 and Fig. 4.).

Parameter β: Parameter β in the Fourier Converter (Sec. 3.2) affects FDRNet at both network training and document recognition (inference) stages. Specifically, FDRNet converges well when β lies within a suitable range. In addition, FDRNet recognition is not sensitive to β either. As Tab. 6 shows, the CER of the trained FDRNet models is quite stable when β changes in certain ranges. More details about parameter β on model training are provided in the Supplementary Material.

Constraints: The proposed FDRNet may be confused if the document background region has similar patterns as the document region. Under such situations, document background could be treated as parts of document region in restoration as illustrated in the last sample in Fig. 6.

5. Conclusion and Future Work

This paper presents a document restoration network FDRNet for better recognition of document images captured by cameras. FDRNet focuses on high-frequency information in the Fourier space which allows it to learn from a small amount of training data effectively. Additionally, FDRNet can generalize to new data well as it discards low-frequency information which mitigates domain gaps greatly. Extensive experiments show that FDRNet is capable of removing geometric and appearance degradation which improves document recognition significantly. In the future, we would like to study the simple yet effective image synthesis and so to leverage the advances of training on both real and synthetic data for more robust document dewarping and recognition.

References

[1] Marcos Almeida, Rafael Dueire Lins, Rodrigo Bernardino, Darlisson Jesus, and Bruno Lima. A new binarization algorithm for historical documents. *Journal of Imaging*, 4(2):27, 2018. 3

[2] Hmrishav Bandyopadhyay, Tanmoy Dasgupta, Nibaran Das, and Mita Nasipuri. A gated and bifurcated stacked u-net module for document image dewarping. *arXiv preprint arXiv:2007.09824*, 2020. 6

[3] Su Bolan, Lu Shijian, and Chew Lim Tan. A self-training learning document binarization framework. In *2010 20th International Conference on Pattern Recognition*, pages 3187–3190. IEEE, 2010. 3

[4] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 11(6):567–585, 1989. 3

[5] Michael S Brown and W Brent Seales. Document restoration using 3d shape: a general deskewing algorithm for arbitrarily warped documents. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 367–374. IEEE, 2001. 2

[6] Frédéric Courteille, Alain Crouzil, Jean-Denis Durou, and Pierre Gurdjos. Shape from shading for the digitization of curved documents. *Machine Vision and Applications*, 18(5):301–316, 2007. 2

[7] Ricardo da Silva Barbosa, Rafael Dueire Lins, Edson Da F De Lira, and Antonio Carlos A Camara. Later added strokes or text-fraud detection in documents written with ballpoint pens. In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 517–522. IEEE, 2014. 3

[8] Sagnik Das, Ke Ma, Zhixin Shu, Dimitris Samaras, and Roy Shilkrot. Dewarpnet: Single-image document unwarping with stacked 3d and 2d regression networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 131–140, 2019. 2, 6, 8

[9] Sagnik Das, Gaurav Mishra, Akshay Sudharshana, and Roy Shilkrot. 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*, pages 125–128, 2017. 2

[10] Hironori Ezaki, Seiichi Uchida, Akira Asano, and Hiroaki Sakoe. Dewarping of document image by global optimization. In *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, pages 302–306. IEEE, 2005. 2

[11] Hao Feng, Yuechen Wang, Wengang Zhou, Jiajun Deng, and Houqiang Li. Doctr: Document image transformer for geometric unwarping and illumination correction. *arXiv preprint arXiv:2110.12942*, 2021. 2, 3, 6

[12] Matteo Frigo and Steven G Johnson. Fftw: An adaptive software architecture for the fft. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 3, pages 1381–1384. IEEE, 1998. 4

[13] Bin Fu, Minghui Wu, Rongfeng Li, Wenxin Li, Zhuoqun Xu, and Chunxu Yang. A model-based book dewarping method using text line detection. In *Proc. 2nd Int. Workshop on Camera Based Document Analysis and Recognition, Curitiba, Barazil*, pages 63–70, 2007. 2

[14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 3

[15] Hyung Il Koo, Jinho Kim, and Nam Ik Cho. Composition of a dewarped and enhanced document image from two view images. *IEEE Transactions on Image Processing*, 18(7):1551–1562, 2009. 2

[16] Xiaoyu Li, Bo Zhang, Jing Liao, and Pedro V Sander. Document rectification and illumination correction using a patch-based cnn. *ACM Transactions on Graphics (TOG)*, 38(6):1–11, 2019. 3

[17] Jian Liang, Daniel DeMenthon, and David Doermann. Geometric rectification of camera-captured document images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):591–605, 2008. 2

[18] Rafael Dueire Lins, Rodrigo Barros Bernardino, Darlisson Marinho de Jesus, and José Mário Oliveira. Binarizing document images acquired with portable cameras. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 6, pages 45–50. IEEE, 2017. 3

[19] Rafael Dueire Lins, Ergina Kavallieratou, Elisa Barney Smith, Rodrigo Barros Bernardino, and Darlisson Marinho de Jesus. Icdar 2019 time-quality binarization competition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1539–1546. IEEE, 2019. 3

[20] Changsong Liu, Yu Zhang, Baokang Wang, and Xiaoqing Ding. Restoring camera-captured distorted document images. *International Journal on Document Analysis and Recognition (IJDAR)*, 18(2):111–124, 2015. 2

[21] Xiyan Liu, Gaofeng Meng, Bin Fan, Shiming Xiang, and Chunhong Pan. Geometric rectification of document images using adversarial gated unwarping network. *Pattern Recognition*, 108:107576, 2020. 6

[22] Shijian Lu, Ben M Chen, and Chi Chung Ko. A partition approach for the restoration of camera images of planar and curled document. *Image and Vision Computing*, 24(8):837–848, 2006. 2

[23] Shijian Lu, Bolan Su, and Chew Lim Tan. Document image binarization using background estimation and stroke edges. *International Journal on Document Analysis and Recognition (IJDAR)*, 13(4):303–314, 2010. 3

[24] Shijian Lu and Chew Lim Tan. Document flattening through grid modeling and regularization. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 1, pages 971–974. IEEE, 2006. 2

[25] Shijian Lu and Chew Lim Tan. The restoration of camera documents through image segmentation. In *International Workshop on Document Analysis Systems*, pages 484–495. Springer, 2006. 2

[26] SJ Lu and Chew Lim Tan. Binarization of badly illuminated document images through shading estimation and compensation. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 1, pages 312–316. IEEE, 2007. 3

[27] Ke Ma, Zhixin Shu, Xue Bai, Jue Wang, and Dimitris Samaras. Docunet: document image unwarping via a stacked u-net. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4709, 2018. 2, 5, 6, 7

[28] Gaofeng Meng, Chunhong Pan, Shiming Xiang, Jiangyong Duan, and Nanning Zheng. Metric rectification of curved document images. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):707–722, 2011. 2

[29] Gaofeng Meng, Yuanqi Su, Ying Wu, Shiming Xiang, and Chunhong Pan. Exploiting vector fields for geometric rectification of distorted document images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–187, 2018. 2

[30] Gaofeng Meng, Ying Wang, Shenquan Qu, Shiming Xiang, and Chunhong Pan. Active flattening of curved document images via two structured beams. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3890–3897, 2014. 2

[31] Jonas Östlund, Aydin Varol, Dat Tien Ngo, and Pascal Fua. Laplacian meshes for monocular 3d shape recovery. In *European conference on computer vision*, pages 412–425. Springer, 2012. 2

[32] René Schuster, Oliver Wasenmuller, Christian Unger, and Didier Stricker. Sdc-stacked dilated convolution: A unified descriptor network for dense matching tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2556–2565, 2019. 4

[33] Ray Smith. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE, 2007. 8

[34] Nikolaos Stamatopoulos, Basilis Gatos, Ioannis Pratikakis, and Stavros J Perantonis. Goal-oriented rectification of camera-based document images. *IEEE transactions on image processing*, 20(4):910–920, 2010. 2

[35] Bolan Su, Shijian Lu, and Chew Lim Tan. Binarization of historical document images using the local maximum and minimum. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pages 159–166, 2010. 3

[36] Bolan Su, Shijian Lu, and Chew Lim Tan. Robust document image binarization technique for degraded document images. *IEEE transactions on image processing*, 22(4):1408–1417, 2012. 3

[37] Yuandong Tian and Srinivasa G Narasimhan. Rectification and 3d reconstruction of curved document images. In *CVPR 2011*, pages 377–384. IEEE, 2011. 2

[38] Yau-Chat Tsoi and Michael S Brown. Multi-view document rectification using boundary. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 2

[39] Adrian Ulges, Christoph H Lampert, and Thomas Breuel. Document capture using stereo vision. In *Proceedings of the 2004 ACM symposium on Document engineering*, pages 198–200, 2004. 2

[40] Adrian Ulges, Christoph H Lampert, and Thomas M Breuel. Document image dewarping using robust estimation of curled text lines. In *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, pages 1001–1005. IEEE, 2005. 2

[41] Toshikazu Wada, Hiroyuki Ukida, and Takashi Matsuyama. Shape from shading with interreflections under a proximal light source: Distortion-free copying of an unfolded book. *International Journal of Computer Vision*, 24(2):125–135, 1997. 2

[42] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 1451–1460. IEEE, 2018. 4

[43] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. IEEE, 2003. 6

[44] Atsushi Yamashita, Atsushi Kawarago, Toru Kaneko, and Kenjiro T Miura. 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, 2004. ICPR 2004.*, volume 1, pages 482–485. IEEE, 2004. 2

[45] Shaodi You, Yasuyuki Matsushita, Sudipta Sinha, Yusuke Bou, and Katsushi Ikeuchi. Multiview rectification of folded documents. *IEEE transactions on pattern analysis and machine intelligence*, 40(2):505–511, 2017. 2, 6

[46] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 4

[47] Li Zhang, Andy M Yip, Michael S Brown, and Chew Lim Tan. A unified framework for document restoration using inpainting and shape-from-shading. *Pattern Recognition*, 42(11):2961–2978, 2009. 2

[48] Li Zhang, Yu Zhang, and Chew Tan. An improved physically-based method for geometric restoration of distorted document images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):728–734, 2008. 2