End-to-end Piece-wise Unwarping of Document Images

Sagnik Das[1,2] Kunwar Yashraj Singh[1] Jon Wu[1] Erhan Bas[1] Vijay Mahadevan[1]
Rahul Bhotika[1] Dimitris Samaras[2]

[1]Amazon AI [2]Stony Brook University

{sinkunwa, jonwu, erhanbas, vmahad, bhotikar}@amazon.com, {sadas, samaras}@cs.stonybrook.edu

Abstract

Document unwarping attempts to undo physical deformations of the paper and recover a 'flatbed' scanned document-image for downstream tasks such as OCR. Current state-of-the-art relies on global unwarping of the document which is not robust to local deformation changes. Moreover, a global unwarping often produces spurious warping artifacts in less warped regions to compensate for severe warps present in other parts of the document. In this paper, we propose the first end-to-end trainable piece-wise unwarping[1] method that predicts local deformation fields and stitches them together with global information to obtain an improved unwarping. The proposed piece-wise formulation results in 4% improvement in terms of multi-scale structural similarity (MS-SSIM) and shows better performance in terms of OCR metrics, character error rate (CER) and word error rate (WER) compared to the state-of-the-art.

1. Introduction

Document images captured using mobile devices often contain artifacts due to the physical shape of the paper, camera pose or complex lighting conditions. Therefore, unlike images captured with high fidelity using flatbed scanners, mobile captured documents are ill-suited for digitization.

To improve the image quality for downstream tasks, such as OCR, document unwarping is used to minimize the visible distortion between a captured document image and its flatbed-scanned version. With multiple sources of distortion due to camera viewpoint, paper shape and illumination, the task of unwarping document images in-the-wild is inherently challenging and a long-standing research problem in the domain of document analysis. Most solutions to date, first estimate the deformed 3D shape, and then unwarp the image to make it planar. There is a large body of work in 3D shape-based unwarping, some relying on specific hardware such as stereo [33, 28] and structured light [3, 23], or using 2D images to estimate 3D using shape-from-shading [30, 35] or multi-view images [27, 34]. With respect to image based document unwarping, traditional approaches detect the boundary of a document [4], or explicitly predict text-lines [26, 21]. Generally, these traditional approaches are not very accurate, e.g. textline based methods work as expected only if there exist enough textlines in

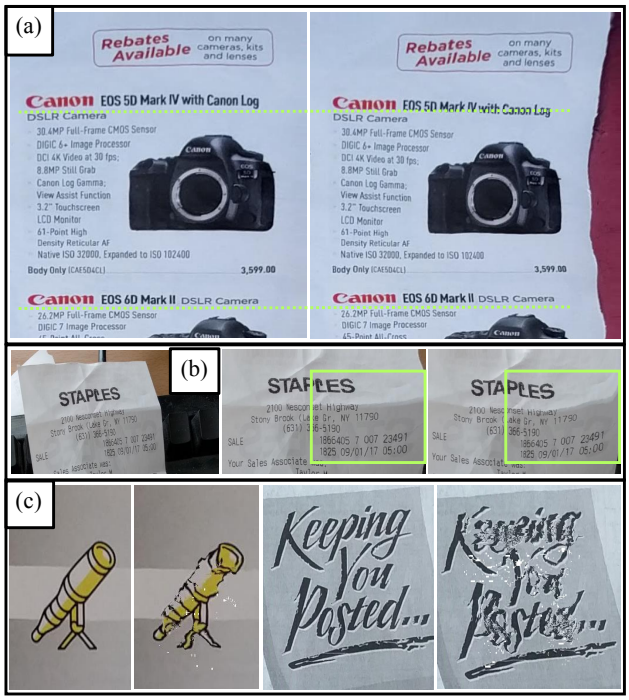

Figure 1. **Comparison to state of the art with the proposed piece-wise approach**: (a) Our method (left) show local improvements, text-lines are more aligned (highlighted with dashed green-line) and leads to better unwarping on the edges compared to [7] on the right. (b) Original image (left), result of the proposed approach (center) and that of global unwarping [7] (right) shows that global unwarping introduces additional warping (highlighted in box). (c) Artifacts due to the absence of stitching reconstruction loss in [14], columns 1, & 3 show our results without such artifacts.

[1]Project page: https://sagniklp.github.io/PiecewiseUnwarp/
This project was initiated when Sagnik Das was an intern at AWS.

the image. With the progress of deep learning, most of the recent methods have become end-to-end and data-driven.

End-to-end deep learning based document unwarping approaches such as DocUNet [19], DewarpNet [7] and CREASE [20] directly predict the global unwarping map. However, these approaches mainly focus on global information and tend to overlook local information. This often results in (1) less robust local unwarping, (2) unexpected warping as shown in Fig. 1 (a) (b). There are few approaches that successfully apply local unwarping be it through using patches [14] or 2D segments [8]. Notably, none of these piece-wise methods are end-to-end trainable, and therefore have difficulty in generalizing well to arbitrary scenarios such as large deformation of the paper. Moreover, optimization based patch stitching used in [14, 8] often lead to undesired stitching artifacts in the output unwarped images (see figure 1c).

Inspired by these facts, in this paper, we propose the first end-to-end trainable piece-wise unwarping approach made possible by a novel fully-differentiable feature-level stitching module for the local unwarping maps. Our goal is to leverage local information for better document unwarping. Specifically, we argue that learning local and global deformation separately through the use of patches, as well as learning the appropriate patch stitching will better utilize local shape deformations. A local approach is also motivated by the fact that a complexly folded/warped document is a combination of multiple simpler deformations which are easier to be approximated locally.

Our approach consists of three trainable modules: (a) shape network (SNet), (b) piece-wise unwarping network (PUNet) and, (c) global stitching network (GSNet). The *SNet* takes the image as input and outputs a 3D shape of the paper. The *PUNet* takes 3D shape patches as input and regresses local unwarping maps. The *GSNet* takes the local patches as input and outputs a global unwarping map to unwarp the input image. All three networks are trained end-to-end with losses on local and global unwarping map regression, and final unwarped image reconstruction. The main contributions of our paper are:

First, a novel end-to-end trainable framework that estimates document unwarping in a piece-wise manner, focusing on unwarping the local deformations.

Second, a fully differentiable stitching network that takes the per-patch unwarping map as input and produces a global unwarping map. This stitching module is end-to-end trainable, and generates artifact free unwarped images, which improves upon prior stitching-based works [14].

Third, we show significant improvement in local unwarping quality, with the proposed piece-wise approach. We improve the prior state-of-the-art in terms of the image similarity metric, MS-SSIM and show more stable performance in terms of OCR error metrics.

2. Related Work

2.1. Non Deep Learning Based Approaches

Parametric model based methods assume the deformation of the document can be represented by low dimensional parametric surface models such as Cylindrical surface [9], Coon's patch [11]. These models are either designed using visual cues such as text lines [21], content-driven vector field [22], boundaries [5, 8] or structured lights [23]. Besides surface models, spline curves are used to model a deformed paper, such as NURBS [33], spline [10], Natural Cubic Splines (NCS) [27].

These low dimension parametric models cannot model very complex surface deformation with multiple folds. This drawback limits their usage to only certain cases, like curls or perspective distortion. However, these models are still useful when the a complex shape can be systematically divided into simpler deformations [8].

Mesh based methods use discrete surface representation for document shape and mainly work in two steps - first estimate a shape then estimate the unwarping. They directly estimate the position of each vertex of the mesh and employ different 3D estimation approaches such as stereo vision [28], point cloud fitting [3], laser scanner [36], shape from texture [17, 29, 26], multiview imaging [34].

Estimating the deformed mesh of the documents relies on well-calibrated and expensive setup of hardware or significant assumptions on document content or multi-view images. With all of these assumptions, the application of the unwarping method becomes limited in realistic scenes.

2.2. Deep Learning Approaches

End-to-end unwarping approaches do not make assumptions about the image texture and do not require any calibrated hardware setup. They are easy to deploy and generalize well on real images. The first deep learning based approach was DocUNet [19], which directly regresses the forward mapping from the deformed document image. However, this method was trained on synthetic images created using random 2D deformations and is therefore unable to exploit the 3D geometric properties of the paper warping and often generates unrealistic results in testing. The successor to this work, DewarpNet [7], regresses the unwarping map using an intermediate 3D shape supervision which improves the generalization in testing due to the disentangled 3D shape representation. A different deep learning based approach, CREASE [20], proposes additional content based loss functions for DewarpNet training. Recently, AGUN [18] proposed a generative adversarial learning based approach for unwarping.

Non end-to-end approaches utilize CNNs to recover the document deformation first and then employ a computational step for unwarping [8, 14]. These methods generally work in a piece-wise fashion. Das et al. [8] utilizes

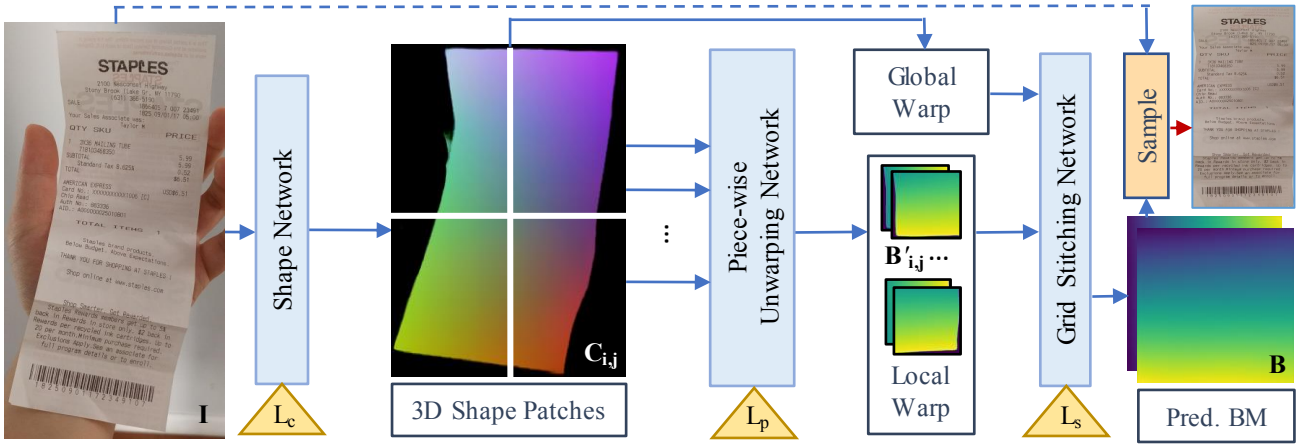

Figure 2. **Proposed Architecture**: The SNet takes an image and produces a 3D shape map, the proposed PUNet takes shape patches as input, and regresses the local backward maps. The global BM is regressed by stitching the local BMs using a novel feature pyramid stitching module, GSNet. An additional global branch is used to guide the stitching network with global alignment and scale information. The triangles denote the loss functions. Test time unwarping step is shown using dashed arrow. Final bilinear sampling step (denoted by 'Sample' block) includes a resize operation and implemented using PyTorch $grid\_sample$ function.

semantic segmentation to detect the fold lines and divide the document in multiple parts. The unwarping for each part is estimated using Coon's Patch [11]. DocProj [14] estimates a per patch vector flow field for unwarping using a deep network then employs a graph-cut based stitching approach. The graph-cut optimization objective is not differentiable and hence not end-to-end trainable. In contrast to [14], our framework is guided by the reconstruction loss on the unwarped image which automatically imposes global constraints during stitching. Additionally, the reconstruction loss reduces stitching artifacts in the unwarped images (see figure 1c).

3. Piece-wise Unwarping

The proposed piece-wise unwarping network is composed of three sub-networks that are designed for 3D shape regression, piece-wise unwarping backward map (BM) regression and stitching of the regressed warping fields. The schematic overview of the proposed approach is shown in figure 2. These networks are accordingly named as the (1) Shape Network (SNet), (2) Piece-wise Unwarping Network (PUNet) and (3) Global Stitching Network (GSNet).

3.1. Shape Network

The goal of the SNet is to transform an input image I to a per pixel 3D coordinate map, $C \in \mathcal{R}^{w \times h \times 3}$, where each pixel value (X, Y, Z) corresponds to the 3D coordinates of the document shape. This representation encodes the 3D shape of the paper and also implicitly encodes the camera projection parameters which is sufficient to learn the backward map (BM) for unwarping. Moreover, the shape representation enables solving the unwarping task in a more physically constrained domain rather than learning from a joint distribution of document texture, shape, illumination and, camera perspective. The design of SNet follows from the 3D shape regression network proposed in [7]. We treat the task as an image-to-image translation problem and use a UNet style encoder-decoder for implementation.

Loss Functions. To train the SNet we utilize the L_1 error between the predicted ($\hat{C}$) and ground-truth (C) 3D coordinate maps. Additionally, we apply an image gradient based loss term on C for better reconstruction of sharp curvature changes, e.g. the folds. The loss function is given by $L_c = ||C - \hat{C}||_1 + \alpha||\nabla C - \nabla \hat{C}||_1$. Here, ∇C denotes the horizontal and vertical gradient of C. α controls the weight of the gradient term.

3.2. Piece-wise Unwarping Network

Learning a global unwarping often leads to sub-optimal result and often times the network introduces undesired warping to lesser warped regions of the document (See figure 1 (b)). With piece-wise unwarping, we provide robustness to local shape variations. To achieve this, the 3D coordinate map (C) and backward map (B) is partitioned into n^2 non-overlapped patches $\{C_{i,j}\}$ and $\{B_{i,j}\}$.

$$C_{i,j} = \left[\frac{i}{n}, \frac{i+1}{n}\right) \times \left(\frac{j}{n}, \frac{j+1}{n}\right], \quad 0 \leq i, j < n \quad (1)$$

Where i and j denotes the row and column index of the patches. The corresponding BM B is partitioned as:

$$B_{i,j} = \begin{cases} [B_u, B_v] & \frac{i}{n} \leq u < \frac{i+1}{n}, \frac{j}{n} \leq v < \frac{j+1}{n} \\ -1 & \text{otherwise} \end{cases} \quad (2)$$

Where $[B_u, B_v]$ is the value of the BM at location (u, v) which contains a canonical pixel coordinate of I to be sampled at (u, v). The -1 values denote the invalid pixels not present in patch $C_{i,j}$.

We normalize each $B_{i,j}$ w.r.t. the canonical coordinate range of the C_{ij} and also tight crop the valid coordinates. We call the normalized and cropped $B_{i,j}$ as the *local BM*, $B'_{i,j}$. For the valid coordinates the normalization operation is given as the following:

$$B'_{i,j} = \left[\frac{B_u - (i/n)}{(i+1)/n - (i/n)}, \frac{B_v - (j/n)}{(j+1)/n - (j/n)} \right] \quad (3)$$

The PUNet takes a patch $C_{i,j} \in \mathcal{R}^{(w/n)\times(h/n)\times 3}$ as input and outputs the local backward map, $B'_{i,j} \in \mathcal{R}^{(w/n)\times(h/n)\times 2}$. This network is implemented as an encoder-decoder of DenseNet blocks with Layer Normalization [2]. Using BatchNorm [12] leads to over-fitting since the C_{ij} in a batch are highly correlated.

Loss Functions. Initially, PUNet is trained with L_1 loss on the local BM, $B'_{i,j}$ and L_2 loss on the predicted unwarped image patches, $D_{i,j}$. After the first round of training with ground-truth $C_{i,j}$ as input, we perform end-to-end training of the PUNet and SNet. For this step we utilize the SNet predicted $\hat{C}_{i,j}$ as inputs to PUNet. The complete loss function to train the PUNet is given as: $L_p = ||B'_{i,j} - \hat{B}'_{i,j}||_1 + \beta_1||D_{i,j} - \hat{D}_{i,j}||_2 + \beta_2 L_c$. Where $D_{i,j}$ and $\hat{D}_{i,j}$ denotes the input image patches unwarped using ground-truth BM patch, $B'_{i,j}$ and predicted BM patch $\hat{B}'_{i,j}$. Following DewarpNet [7] we use the checkerboard images for $D_{i,j}$ (more details are discussed in supplementary). For the PUNet initial training with $C_{i,j}$, β_2 is set to 0.

3.3. Global Stitching Network

We propose a feature level stitching network to stitch the local BMs to regress B which is used to unwarp the image I. Although, it is possible to design the stitching at the image level using image registration strategies, it'll be susceptible to misalignment and ghosting [8] mainly due to texture-less regions in a document. Conversely, a feature level approach on the local BMs is robust to the aforementioned problems. Also, since the local BMs differ in scale due to the perspective distortion present in I we employ a feature pyramid in the stitching network. Finally, to ensure better global alignment we introduce a global BM feature branch as a residual to the local BM pyramidal features.

The proposed global stitching network (GSNet) consists of two sub-modules, (1) Canonical Placement Module (CPM) and (2) Grid Stitching Feature Pyramid Network (G-FPN). An overview of GSNet is shown in figure 5.

Canonical Placement Module. After the normalization step in eq. 3, each valid local BM position $[B_u, B_v] \in [0, 1]$ encodes canonical coordinates of the corresponding image patch, $I_{i,j}$. In order to unwarp the image I, $[B_u, B_v]$ values need to be rescaled to the canonical coordinates of I. Therefore prior to the stitching, the local BMs $B'_{i,j}$ are denormalized using the inverse operation of eq. 3.

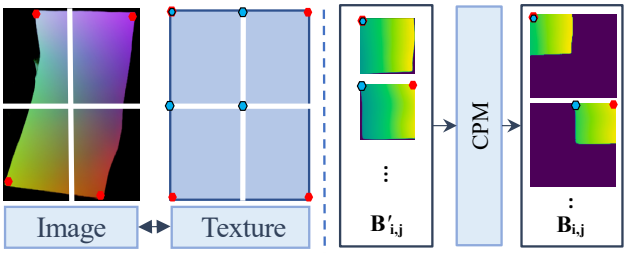

Figure 3. **Canonical Placement of Local BMs.** *Left:* Red markers show the corner correspondence between the image and the texture (unwarped) domain assuming the document is a equiangular quadrilateral. The blue markers show the patch origins $B\left(\frac{i}{n}, \frac{j}{n}\right)$ utilized in CPM. *Right:* CPM module: **red** shows the corner, and **blue** shows the patch origins utilized for global coarse placement of the local patches.

After the denormalization step, we perform a coarse spatial placement of each local BM. Assuming a document is a quadrilateral we can assume the top leftmost image patch, $C_{0,0}$ will unwarp to the top leftmost part of B and so on. We show an illustration in 3. By exploiting this spatial correspondence of $C_{i,j}$ and $B'_{i,j}$, each local BM is placed at $B\left(\frac{i}{n}, \frac{j}{n}\right)$. We can demonstrate a simplest case with $n = 2$, where $B_{0,0}$ is placed at $B(0, 0)$, $B_{0,1}$ at $B(0.5, 0)$, $B_{1,0}$ at $B(0, 0.5)$ and $B_{1,1}$ at $B(0.5, 0.5)$. Illustration of the unwarped patches after the coarse placement step is shown in figure 4 (bottom row). The denormalization and coarse placement step eases the task of the G-FPN by roughly aligning the input $[B_u, B_v]$ values with the output B.

Grid Stitching Feature Pyramid Network. The output local BMs are in different scales due to the perspective difference of the image patches, e.g. a patch closer to the camera has higher scale than a patch far away from camera. Consequently, an unwarped patch closer to the camera has higher scale than a patch further away. This is illustrated in the figure 4 (bottom row). We handle this scale mismatch by employing a feature pyramid while stitching. To perform stitching in the feature space we propose a novel feature pyramid encoder [15] based on Residual Channel Attention Network (RCAN) [37] blocks, initially introduced for image super-resolution. Our stitching task is analogous to super-resolution in the sense that we aim to preserve high-frequency details of the learned local BMs. We use stride-s convolutions to reduce the spatial resolution of local BM features by factor s at each feature level. An overview of G-FPN is illustrated in figure 5. The primary input to G-FPN is the concatenation of the n^2 local BMs obtained from the CPM. To assist the G-FPN with a consistent global scale and alignment, an additional global branch is introduced following [7]. This branch takes the 3D coordinate map $C \in \mathcal{R}^{h\times w\times 3}$ as input and learns to regress B. The output of the global branch is used as a secondary input to the G-FPN. The extracted local features from each level of the pyramid is concate-

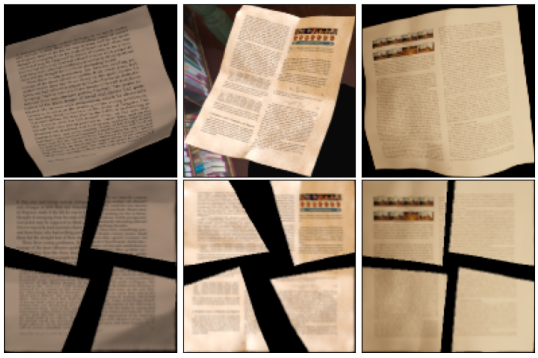

Figure 4. **Illustration of the scale mismatch among different patches:** Top is the input image, bottom is the unwarped patches after the spatial placement in the texture space using the CPM.

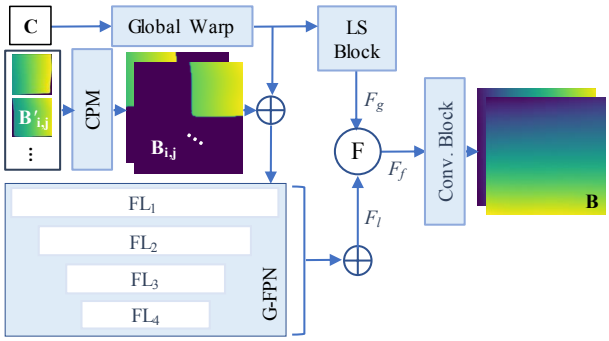

Figure 5. **Proposed GSNet:** CPM denote Canonical Placement Module, Global Warp denote the global branch for global BM, LS block denote a few few conv layers with global features, **F** denote feature fusion ($\mathcal{F}$) for local (F_l) and global (F_g) features, FL_i denote the pyramid level i. $\oplus$ denotes channel-wise concatenation.

nated (F_l) and fed to a global fusion block ($\mathcal{F}$) (also denoted by **F** in figure 5) along with the global features (F_g) extracted from the global branch. An ablation experiment shows the effect of the global and local branch in Table 2. $F_l = \bigoplus_{i=1}^{n} f_i$; $F_f = \mathcal{F}(F_l + F_g)$; $B = \mathcal{F}_t(F_f)$. Here f_i denotes the local features extracted from each feature level i and $\oplus$ denotes the channel-wise concatenation operation. F_f is the output of the global fusion block. $\mathcal{F}_t$ denotes the final block which takes the fused features and outputs B.

In summary, input to the global stitching module is concatenation of the local BMs and the global BM, $\tilde{B} \in \mathcal{R}^{h \times w \times 2(n^2+1)}$ and output is the BM, $B \in \mathcal{R}^{h \times w \times 2}$. We show a comparative evaluation of different stitching model variants in supplementary material. These modules vary in terms of the global fusion function ($\mathcal{F}$) and the long skip block (LS block) structure.

Loss Functions. To train the PUNet we utilize the losses on the final BM and the unwarped image. The respective loss functions are given as: $L_s = ||B - \hat{B}||_1 + \lambda||D - \hat{D}||_2$. $\hat{B}$ denote the stitched BM resulting from G-FPN and B is the respective ground-truth. $\hat{D}$ and D are the input images unwarped using $\hat{B}$ and B respectively. λ denote the weight associated with the second term of the L_s. Similar to [7] we utilize the checkerboard pattern images to obtain D and $\hat{D}$.

3.4. Training Details

Dataset. To train our network we use the Doc3D dataset [7]. Doc3D contains 100K synthetic document images rendered using Blender [1]. This dataset utilizes a large set of document textures for rendering and contains 3D shape, C and BM, B as ground-truth. We use a 88K and 8K split for training and validation of our network. Since all our training data is synthetic, and per-module ground truth is available, we follow the common practice in cascaded networks [31] of pre-training each module, to stabilize overall training before performing end-to-end training.

Augmentations. For SNet training we apply random brightness, contrast, hue, saturation shift to the input images with probability 0.5. Additionally we randomly replace the background of these images using random textures from DTD [6] dataset. For PUNet, we apply augmentation by varying the shape patch size in $[0.4, 0.6]$. We also use variable padding around the C. Without these augmentation steps the PUNet becomes biased and fails to handle padding variability at test time. We also noticed that it is likely that a only small region of a document is visible within a patch ($C_{i,j}$), which destabilizes the training. We utilize minimum bounding rectangle around the document mask to homography transform the image during the training of PUNet.

Hyperparameters. SNet is trained with 256×256 sized images. For PUNet we set $n = 2$ and input 128×128 sized shape patches $C_{i,j}$. PUNet outputs same sized local BM predictions. Each local BM is then resized to $128/n$ and used as an input to CPM. Outputs of CPM and inputs to the global texture stitching module is 128×128. We use 5 Residual Channel Attention Blocks [37] to construct the feature pyramid network, and use 4 times feature reduction in the channel attention block. To train each network, we use Adam [13] optimizer with an initial learning rate of $1e-5$. The learning rate is halved if the validation error doesn't decrease in consecutive 5 epochs. SNet and PUNet are first separately trained to convergence using ground-truth and the learned weights are used to initialize the joint training. We set the loss weights $\alpha = 0.5$, $\beta_1 = 0.03$ and $\beta_2 = 0.5$. Similarly the G-FPN is first trained with local BMs obtained from B. Later, we freeze the weights of SNet and PUNet with the best models and fine-tune G-FPN with $\gamma = 0.03$. We found that using higher values for β_1, γ results noises on the unwarped image during testing.

4. Experimental Evaluation

We validated our proposed piece-wise unwarping approach with multiple experiments. We first evaluate our method against current state-of-the-art and then we present

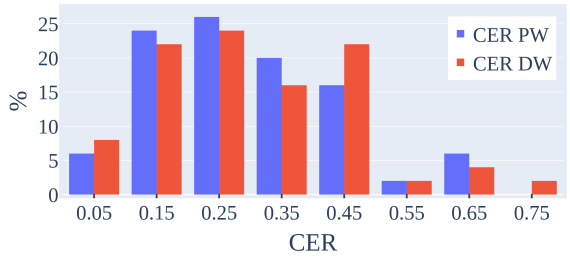

Figure 6. **Distribution of the CER in DocUNet Benchmark**: Proposed (CER-PW) vs. DewarpNet (CER-DW). Higher percentage of documents (y-axis) with lower CER (x-axis) is better.

additional ablation experiments to evaluate our design choices. To evaluate the proposed approach we use the image similarity metric, multi-scale structural similarity (MS-SSIM) and Local Distortion (LD) on the 130-images benchmark from DocUNet [19]. Furthermore, we evaluate OCR performance on an 51-images benchmark of DocUNet using character (CER) and word error-rate (WER).

4.1. Evaluation Metrics

The choice of evaluation metrics follows from the prior document unwarping approaches [7, 19]. The most recent paper CREASE [20] evaluates their method in terms of Edit Distance (ED) [24] between the detected text bounding boxes in ground-truth (scanned) and predicted unwarped images. However, this evaluation scheme requires the ground-truth warping field of the test images and is therefore not applicable for real benchmarks [19]. In this paper, we focus our evaluation on metrics that are applicable to real images. Image similarity metric, MS-SSIM [32] is based on local image statistics (mean and variance) of the unwarped and scanned (ground-truth) images calculated over multiple Gaussian pyramid scales. LD is based on the dense SIFT flow [16] between the unwarped and scanned images. Details about the parameters settings of these evaluation metrics, MS-SSIM, LD, CER and WER are discussed in [7, 19]. We use the same settings for fair comparison. For OCR evaluation we use the open source Tesseract (4.1.1) [25] with the LSTM based OCR engine.

4.2. Comparison with Prior Methods

We quantitatively and qualitatively compare our method with recent deep learning based document unwarping approaches, DocUNet [19], DocProj [14], DewarpNet [7], CREASE [20], and AGUN [18]. DocUNet [19] uses synthetic data and utilizes per-pixel forward mapping regression to learn unwarping in an end-to-end manner. DocProj [14] uses a patch-based approach for local regression with global latent features. But is not end-to-end trainable due to a graph-cut based stitching of the forward map. DewarpNet and CREASE are both trained using the Doc3D dataset, which contains 3D shape of the documents. Our approach differs substantially from DewarpNet as it accounts for, and corrects, deformation at a local patch-level unlike DewarpNet which only performs global unwarping. This is achieved by the local branch consisting of the proposed modules PUNet, CPM, and GSNet. AGUN [18] proposes an adversarial learning based framework to learn document unwarping using synthetically deformed documents in 2D.

Method	MS-SSIM ↑	LD ↓	CER ↓	WER ↓
DocUNet [19]	0.4389	10.90	0.3203 (0.15)	0.4567 (0.20)
DocProj [14]	0.3832	12.83	0.3474 (0.16)	0.4889 (0.21)
AGUN* [18]	0.4491	12.06	-	-
DewarpNet [20]	0.4692	**8.98**	0.3028 (0.16)	0.4368 (0.21)
Proposed	**0.4879**	9.23	**0.3001 (0.14)**	**0.4302 (0.18)**

Table 1. Quantitative comparison of the proposed and prior approaches on DocUNet benchmark dataset. ↑ and ↓ signifies higher and lower better respectively. Standard deviation is reported in the parentheses. *OCR metrics could not be calculated as images or models are not publicly available.

Quantitative Comparison. We compare the proposed piece-wise approach with prior document unwarping approaches, DewarpNet [7], AGUN [18], DocProj [14] and DocUNet [19]. We exclude CREASE from this comparison because their models are not publicly available for a fair quantitative comparison. See Table 1 for the comparison of quantitative results. The proposed piece-wise method outperforms the state-of-the-art DewarpNet in terms of MS-SSIM metric and also shows a small improvement in OCR accuracy. Image similarity based improvement is due to the better local structural alignment of the scanned ground-truth and unwarped images. Although our OCR numbers are very close to DewarpNet, with the proposed piece-wise approach we achieve a lower standard deviation (a 2% reduction), in terms of both CER and WER metrics. To demonstrate the improvement in OCR metrics we show the histogram plot of the CER of all the OCR test documents in figure 6. We can clearly see more documents have a lower CER with the proposed approach. We also show a qualitative OCR error comparison with [7] in figure 10. We must note that the OCR error rates also depend on the accuracy of the OCR engine, and we discuss a few cases in the supplementary where we notice spurious ED values, although the the images are very similar in terms of unwarping quality. Our improvement compared to DocProj [14] is more significant because of two main reasons: (1) [14] assumes local patches have no background, and (2) [14] doesn't use the reconstruction loss during stitching. We would like to highlight that the background assumption is often violated in real images present in the benchmark.

Our method shows a small percentage increase in the LD metric due to global misalignment and scale mismatch with the introduction of local branch. However, these errors are insignificant for document unwarping quality. We discuss our observations in detail in the supplementary material.

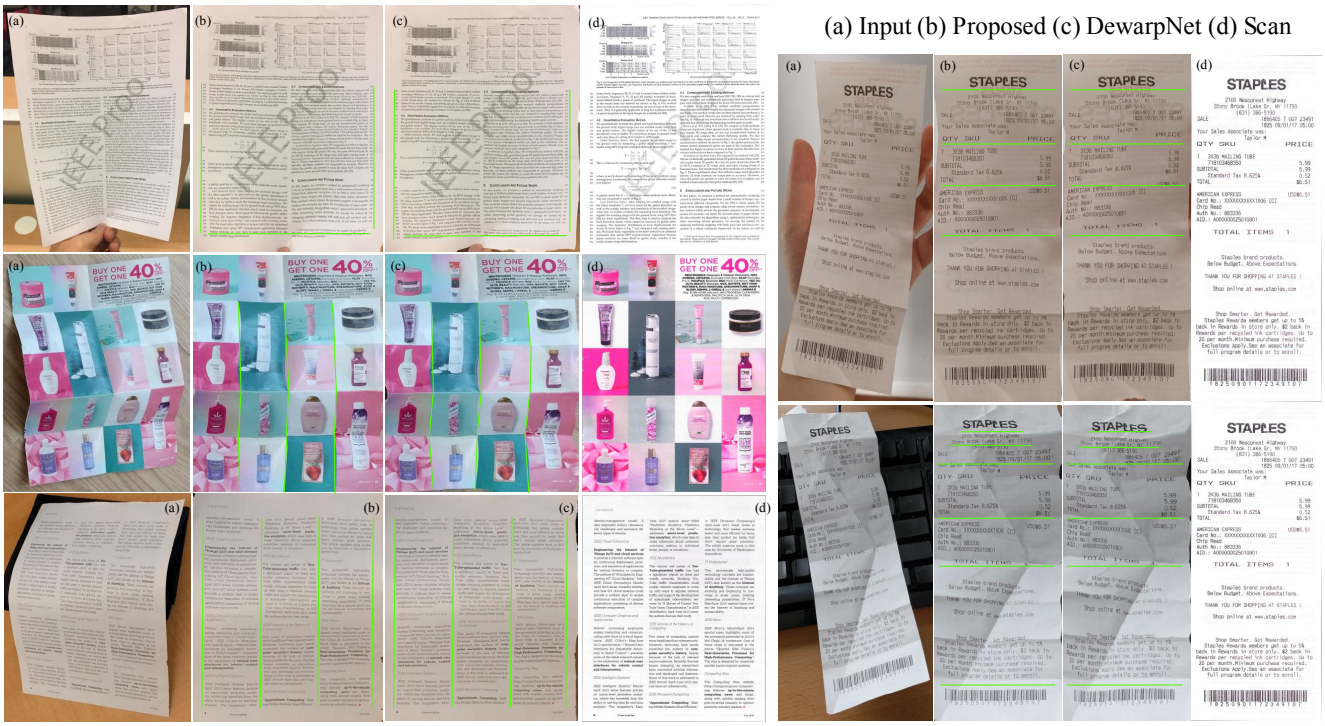

Figure 7. **Qualitative comparison of the piece-wise unwarping and DewarpNet**: (a) input, (b) proposed, (b) DewarpNet [7] , (d) scan gt. The highlighted lines clearly show local improvements.

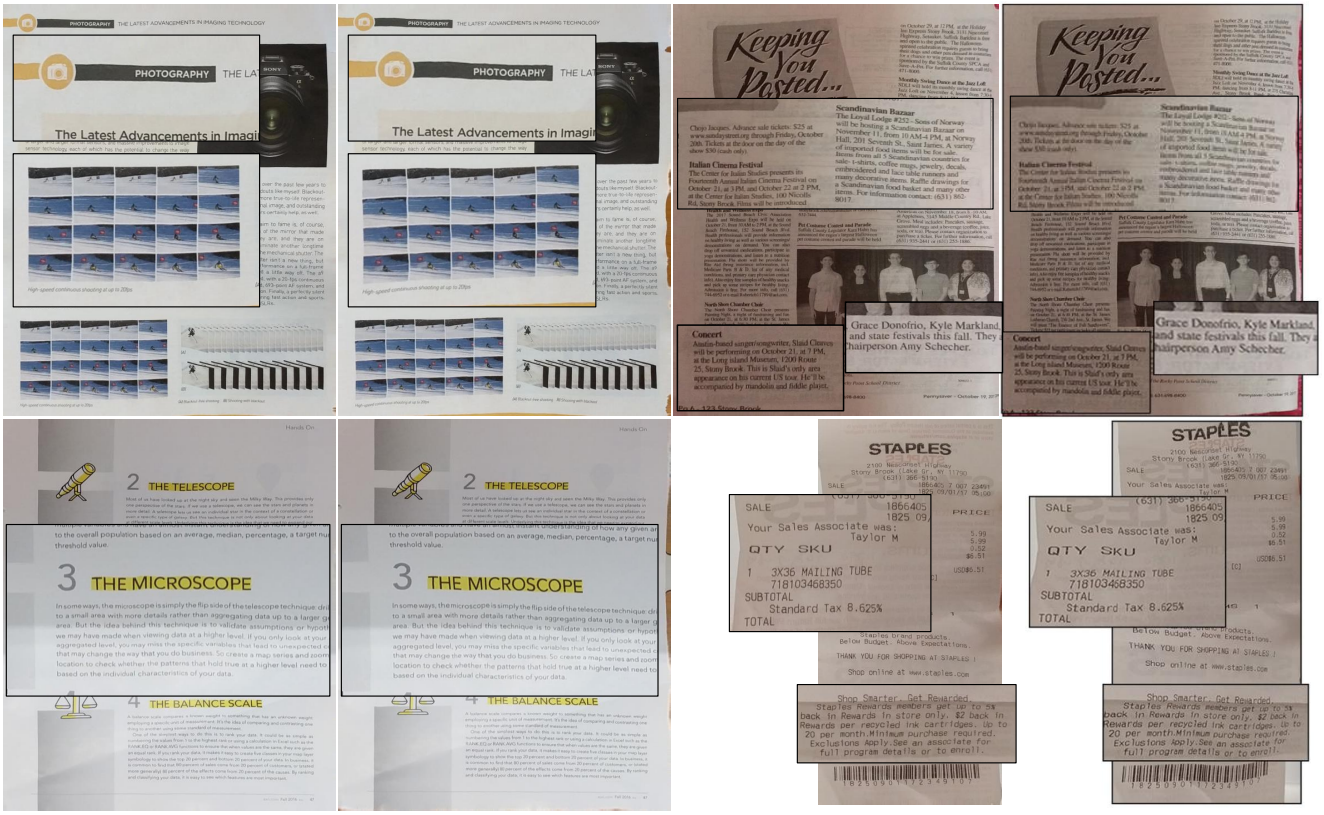

Figure 8. **Local comparison of the proposed method with DewarpNet and CREASE**: Column 1 and 3, shows our results, Column 2 is DewarpNet [7], Column 4 is CREASE [20]. Higher-resolution unwarped images are unavailable for CREASE.

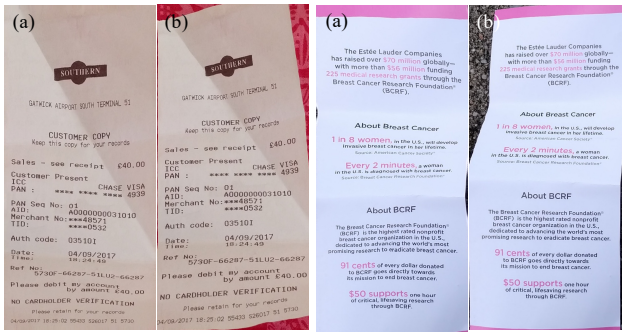

Figure 9. Qualitative comparison of the (a) proposed method and (b) DocProj [14]. [14] assumes local patches have no background, thus fail to handle large background regions in input image.

Figure 10. Qualitative comparison of the OCR on (left) proposed unwarped, (middle) DewarpNet [7] unwarped, and (right) Scanned. We highlight the OCR errors with *red*, number of recognition errors is given in the *yellow* box. Zoom for detail.

Qualitative Comparison. We show a comparison of the qualitative results with DewarpNet in figure 7. To better demonstrate the local improvement due to the the piece-wise approach we show close-up comparisons with global unwarping approaches, DewarpNet [7] and CREASE [20] in figure 8. We can clearly notice that our piece-wise formulation captures local structures, such as text-lines, image boundaries and text segments, better than global strategies. Some of the improvements are highlighted with horizontal and vertical cue lines in figure 7. We also show a qualitative comparison with DocProj [14] in figure 9 which is a patch-wise unwarping approach but not end-to-end differentiable.

4.3. Ablation Studies

This section details the design decisions and ablation studies for our architecture. We would like to note, in addition to the sections below, we have included additional experiments in the supplemental.

Comparison of Global and Local BM Branch. In this experiment, we aim to evaluate the contribution of the global and the local BMs in the global texture stitching module. To separately evaluate each module, we train a stitching FPN without the global BM as input. In this case, only the local BM patches are used as input to the grid stitching FPN to synthesize the final BM. A quantitative comparison of the

Method	MS-SSIM ↑	LD ↓
Local	0.4552	9.78
Global	0.4635	**8.89**
Local+Global	**0.4879**	9.23

Table 2. Comparison of the piece-wise unwarping modules. We evaluate models with only Local or Global branches, and Both.

local stitched BM, global BM and the combination of these two is reported in Table 2. The global branch (the setting corresponding to DewarpNet) shows lower LD than other variants since it achieves better global alignment with the ground truth. On the other hand the local branch independently suffers from global alignment and scale mismatches in cases of large perspective difference between patches as demonstrated in figure 4. The combined module compensates these errors and achieves a 5% higher MS-SSIM over the independent global branch based unwarping.

Performance trade-off between global and local branch. We obtain the best unwarping result by training GSNet with frozen, jointly-trained SNet and PUNet modules. When this constraint is relaxed and SNet, PUNet, GSNet are trained end-to-end, we observe that the GSNet biases the network to focus on global unwarping rather than the local unwarping. In validation, this causes a $\sim$2% L1 error increase in PUNet, and a $\sim$1.6% L1 error decrease in GSNet, thus decreases the overall performance in testing.

Effect of patch overlap. Overlapping between the patches is a common design choice for patch based approaches [14]. However, we don't see any notable improvement with overlapping patches. Our stitching network is trained with global reconstruction loss in an end-to-end manner. It provides a sufficient learning signal so that our model can interpolate the patches even in the absence of any overlapping context. As an advantage of non-overlapping, we do not need to transmit redundant information, and can save on inference costs in memory and runtime.

5. Conclusions and Future Work

We presented a novel end-to-end architecture for piece-wise unwarping of document images. In terms of the image similarity and the OCR metrics we have shown superior performance to prior state-of-the-art approaches. We explicitly model the unwarping as a combination of local and global warping fields, leading to better local reconstruction. For future work, adaptive patching strategies can be leveraged to better incorporate local 3D information. Further, local reconstruction of the stitching network could be extended to a recurrent structure to handle an arbitrary number of patches.

Acknowledgements

This work was partially supported by the Partner University Fund, the SUNY2020 ITSC, and a gift from Amazon AI.

References

[1] Blender - a 3D modelling and rendering package. 5

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[3] Michael S Brown and W Brent Seales. Document restoration using 3D shape: A general deskewing algorithm for arbitrarily warped documents. In *Int. Conf. Comput. Vis.*, 2001. 1, 2

[4] Michael S Brown and Y-C Tsoi. Geometric and shading correction for images of printed materials using boundary. *IEEE Transactions on Image Processing*, 15(6):1544–1554, 2006. 1

[5] Huaigu Cao, Xiaoqing Ding, and Changsong Liu. A cylindrical surface model to rectify the bound document image. In *Int. Conf. Comput. Vis.*, 2003. 2

[6] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014. 5

[7] Sagnik Das, Ke Ma, Zhixin Shu, Dimitris Samaras, and Roy Shilkrot. DewarpNet: Single-image document unwarping with stacked 3D and 2D regression networks. In *Int. Conf. Comput. Vis.*, 2019. 1, 2, 3, 4, 5, 6, 7, 8

[8] 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*, DocEng '17, pages 125–128, 2017. 2, 4

[9] Andrei Doncescu, Alain Bouju, and Veronique Quillet. Former books digital processing: image warping. In *Proceedings Workshop on Document Image Analysis (DIA'97)*, pages 5–9. IEEE, 1997. 2

[10] Hironori Ezaki, Seiichi Uchida, Akira Asano, and Hiroaki Sakoe. Dewarping of document image by global optimization. 2005. 2

[11] Gerald Farin and Dianne Hansford. Discrete coons patches. *Computer aided geometric design*, 16(7):691–700, 1999. 2, 3

[12] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 4

[13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 5

[14] 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)*, 2019. 1, 2, 3, 6, 8

[15] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 4

[16] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):978–994, 2011. 6

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

[18] 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. 2, 6

[19] Ke Ma, Zhixin Shu, Xue Bai, Jue Wang, and Dimitris Samaras. DocUNet: Document Image Unwarping via A Stacked U-Net. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2, 6

[20] Amir Markovitz, Inbal Lavi, Or Perel, Shai Mazor, and Roee Litman. Can you read me now? content aware rectification using angle supervision. *arXiv preprint arXiv:2008.02231*, 2020. 2, 6, 7, 8

[21] Gaofeng Meng, Chunhong Pan, Shiming Xiang, and Jiangyong Duan. Metric rectification of curved document images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):707–722, 2012. 1, 2

[22] 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

[23] Gaofeng Meng, Ying Wang, Shenquan Qu, Shiming Xiang, and Chunhong Pan. Active flattening of curved document images via two structured beams. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014. 1, 2

[24] Frederic P. Miller, Agnes F. Vandome, and John McBrewster. *Levenshtein Distance: Information Theory, Computer Science, String (Computer Science), String Metric, Damerau?Levenshtein Distance, Spell Checker, Hamming Distance*. Alpha Press, 2009. 6

[25] R. Smith. An Overview of the Tesseract OCR Engine. In *ICDAR*. IEEE, 2007. 6

[26] Yuandong Tian and Srinivasa G Narasimhan. Rectification and 3D reconstruction of curved document images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2011. 1, 2

[27] Yau-Chat Tsoi and Michael S Brown. Multi-view document rectification using boundary. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2007. 1, 2

[28] Adrian Ulges, Christoph H. Lampert, and Thomas Breuel. Document Capture Using Stereo Vision. In *Proceedings of the 2004 ACM Symposium on Document Engineering*, DocEng '04, pages 198–200, 2004. 1, 2

[29] Adrian Ulges, Christoph H Lampert, and Thomas M Breuel. Document image dewarping using robust estimation of curled text lines. 2005. 2

[30] 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. 1

[31] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 5

[32] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, 2003. 6

[33] 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 *Int. Conf. Pattern Recog.*, 2004. 1, 2

[34] Shaodi You, Yasuyuki Matsushita, Sudipta Sinha, Yusuke Bou, and Katsushi Ikeuchi. Multiview Rectification of Folded Documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 1, 2

[35] Li Zhang, A. M. Yip, M. 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. 1

[36] 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

[37] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 4, 5