DocMatcher: Document Image Dewarping via Structural and Textual Line Matching

Felix Hertlein
FZI and KIT
Karlsruhe, Germany
hertlein@fzi.de

Alexander Naumann
FZI and KIT
Karlsruhe, Germany
anaumann@fzi.de

York Sure-Vetter
FZI and KIT
Karlsruhe, Germany
york.sure-vetter@fzi.de

Abstract

Document image dewarping is a crucial step in the digitization of physical documents, as it aims to remove the distortions induced by challenging environment settings and document sheet deformations often encountered when using smartphone cameras for image capture. Recently, deep learning-based methods were combined with knowledge about the expected document structure, also known as a template, at inference time to improve the dewarping results. Our contributions in this work are threefold: (1) we propose a novel document image dewarping approach that leverages the prior knowledge about the document structure effectively by detecting and matching lines from the warped and the template domain, and (2) we introduce a novel evaluation metric called matched normalized character error rate (mnCER) to overcome the limitations of existing metrics in evaluating the dewarping process. (3) Finally, we evaluate our approach on the Inv3DReal dataset and show that our approach outperforms the state-of-the-art methods in terms of visual and text-based metrics. Our approach improves upon the state-of-the-art methods by 32.6% in Local Distortion and 40.2% in mnCER. Our code and models are available at https://felixhertlein.github.io/doc-matcher.

1. Introduction

For many business processes, structured documents like invoices, receipts, forms, etc. are a fundamental element of the workflow. Despite the ongoing efforts to digitalize the world, there is still a vast amount of paper documents that need to be processed. The digitization process often starts with capturing the document image using a smartphone or a scanner. Since the latter is less flexible, the trend to use smartphones as a scanning device is increasing. While smartphones are convenient, they introduce new challenges due to the uncontrolled environment settings and a manifold of potential paper deformations (e.g. fold, crumples, creases, etc.). Thus, the captured document images are often distorted, which can lead to poor performance in subsequent document analysis tasks, such as text recognition, information extraction, and visual question answering. To address this issue, document image dewarping aims to remove the distortions induced by the photo-capturing process despite the challenging environment settings.

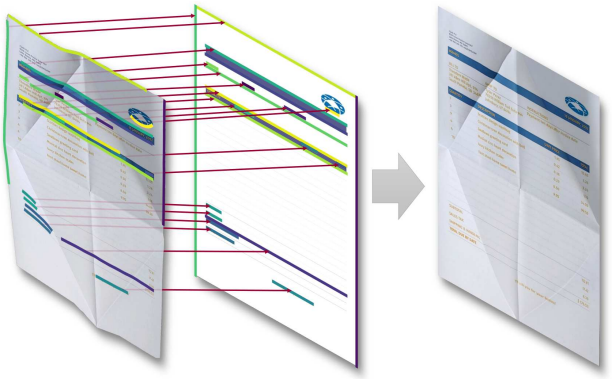


Figure 1. Illustration demonstrating the idea of line detection and matching for document image dewarping.

Recently, the research community has made significant progress in document image dewarping by leveraging deep learning-based methods. Despite promising results, these methods still face challenges when dealing with complex document deformations in difficult environment settings, leaving room for improvement in addressing practical scenarios. Most works consider the document image dewarping as a mapping from a source image to a target image. In a recent publication, the idea of adding additional information about the expected document structure, referred to as a template, at inference time was proposed to improve the dewarping results. This approach has shown promising results, as it allows the model to leverage prior knowledge about the document structure and its visual appearance. It should be noted, that the additional requirement of the tem-

plate image is only a small limitation, as it is often available in practice. For example, in logistics, companies usually use the same invoice layout for all invoices. The receiving company needs to prepare the template just once per supplier and can then use it for all incoming invoices. Digitizing an invoice is then as easy as selecting the invoice type on a mobile device and taking a picture. Importantly, the digitization can be done in the field as the method does not require stationary hardware.

In this work, we propose a novel document image dewarping approach that leverages prior knowledge about the document structure effectively by detecting and matching lines from the warped and the template domain. We use the knowledge of the line correspondences to construct a transformation map that shifts the document elements back to their original position according to the template. This way, we explicitly integrate the template information into the dewarping process and thus, facilitate the downstream document analysis tasks. Figure 1 shows the idea of detecting and matching lines from both domains.

Our contributions are threefold: (1) we propose a novel document image dewarping approach that leverages the prior knowledge about the document structure effectively by detecting and matching lines from the warped and the template domain, and (2) we introduce a novel evaluation metric called mnCER to overcome the limitations of existing metrics in evaluating the dewarping results. (3) Finally, we evaluate our approach on the Inv3DReal [12] dataset and show that our approach outperforms the state-of-the-art methods in terms of visual and text-based metrics.

2. Related Work

Document image dewarping is a well-studied problem in the field of document analysis. The goal is to remove the geometric distortions from the document image to facilitate the subsequent document analysis tasks. In related work, the geometric dewarping is often combined with document illumination correction, namely the removal of lighting and shading effects. In this work, we focus on the geometric dewarping task only, since the illumination correction process can be performed independently.

Template-free Dewarping. In a pioneering work, Ma *et al*. [27] proposed DocUNet, a combination of U-Net architecture to estimate the forward map first and then the backward map. Many works have followed the approach of chaining multiple networks with intermediate supervision [4, 28, 30, 34]. With the emerging transformer architectures, the document dewarping task has been tackled with transformer-based models [8, 41]. While most approaches process the whole document image at once, some works have proposed to divide the document image into patches and process them separately to decouple global and local deformations [5, 21, 22]. Liu *et al*. [25] and Feng *et al*. [7] proposed hierarchical approaches to handle the deformations at different scales. Another approach to reducing the overall distortion is to pre-unwarp the document before applying the dewarping model [3, 24, 35, 37]. Often, the pre-unwarping is done based on the detected document outline, thus, depending strongly on the quality of the outline detection. Our method, however, involves pre-unwarping the image by considering all detected structural and textual lines collectively, making it more robust. Other works have proposed iterative approaches to refine the dewarping results [9, 38, 39]. Thereby, they can handle the global and local deformations with the same model but at different stages of the iteration process. Since most documents contain a considerable amount of text, there have been several works that integrate the text into the dewarping process [2, 10, 13, 20, 26].

Template-based Dewarping. Recently, the idea of using additional information about the expected document structure and appearance has been proposed to improve the geometric dewarping [12] and illumination correction [11]. Hertlein *et al*. [12] introduce the Inv3D dataset, which consists of synthetically generated document images with varying distortions and a corresponding template image. They present a template-based document dewarping model called GeoTrTemplate, which leverages the template information by encoding the template image and the warped image into deep embeddings and connecting them with an attention mechanism. The work does not employ document segmentation, pre-unwarping, nor does it explicitly handle text lines. Our approach builds upon the idea of leveraging the template information but more explicitly by detecting and matching lines from the warped and the template domain.

3. Approach

Our approach consists of several stages, as illustrated in Figure 2. First, the background is removed from the input image. Subsequently, we detect the structural and textual lines in the warped document image and use them to pre-unwarp the document using a homography transformation. We then employ the established document dewarping model GeoTrTemplateLarge [12] twice to initially dewarp the document before we apply our new line-based unwarping method. This method detects structural and textual lines in the pre-unwarped document and matches them to the template lines. Given these line correspondences, we construct a dense transformation map and apply it to the pre-unwarped document to obtain the final dewarped document. We describe each stage in detail below.

3.1. Document Detector

Given a warped document image $\mathbf{W}_0 \in \mathbb{R}^{h \times w \times 3}$, the first step is to remove the background from the image. Thereby, the subsequent steps do not have to deal with the

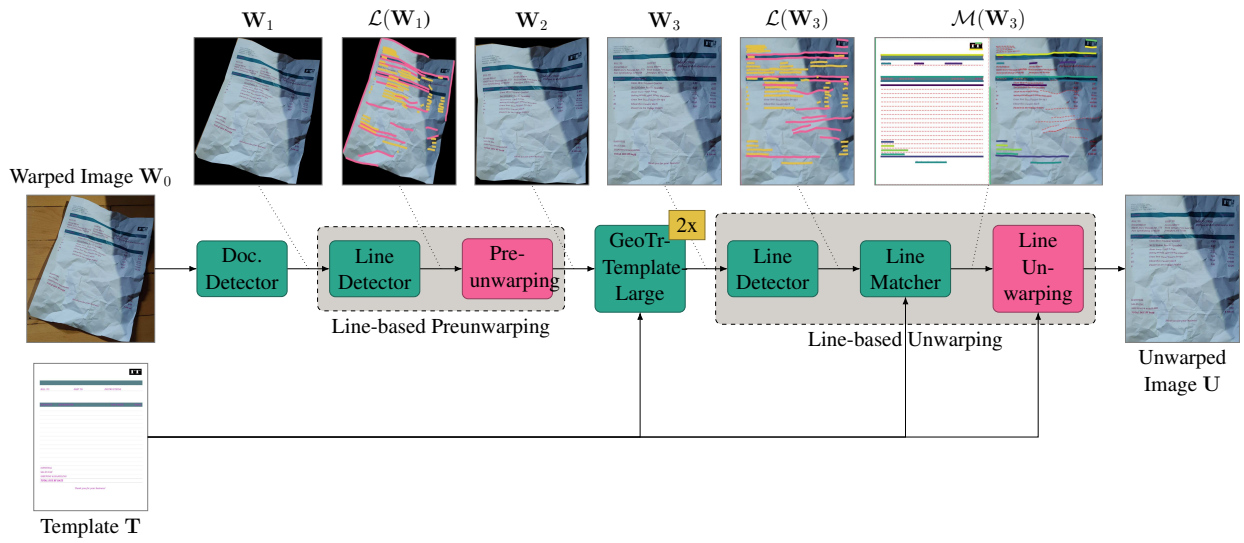


Figure 2. Overview of the proposed approach. First, the background is removed from the input image. Then, the structural and text lines are extracted from the background-removed image. We use the detected lines to pre-unwarp the image and then match the detected lines to the template lines. Given the line matches, we compute the dense transformation field and apply it to the input image.

noise in the background. We achieve this by fine-tuning the state-of-the-art segmentation model Segment Anything (SAM) [14] on Inv3D [12] for the task of semantic segmentation. We set a total of two classes: the document class and the background class. For fine-tuning, we freeze the image and prompt encoder of SAM and only train the mask decoder of the `ViT-L` model. The image input resolution is set to 1024 $\times$ 1024 pixels. The output of the model is a binary mask that segments the document from the background. We then apply the mask to the input image to remove the background. Let $\mathbf{W}_1 \in \mathbb{R}^{h \times w \times 3}$ represent the resulting image with the background removed.

3.2. Line Detector

Our line detector follows the approach of Lal *et al*. [16]. They treat the line detection task as an instance segmentation problem and propose a transformer-based model using masked attention called LineFormer. We train our own version of the LineFormer model on the Inv3D dataset for the task of structural and textual line detection. Since the Inv3D dataset does not provide structural line annotations, we generate the line annotations based on the optimal image $\hat{\mathbf{I}}$ using the Canny edge detector [1]. In contrast to the original LineFormer model, we extract 2D line paths instead of mathematical functions. In a post-processing step, we remove duplicate lines and lines with a length below a certain threshold. We denote the detected lines as $\mathcal{L}(\mathbf{W}_i)$ for $\mathbf{W}_i$ being the input image.

3.3. Pre-Unwarping

To simplify the problem, we pre-unwarp the image using the detected lines by estimating a homography transformation. We base the pre-unwarping on the detected structural and textual lines instead of the document outline, as the document outline is not always correctly detected in hard cases. The idea behind our homography estimation is to axis align all detected lines as well as possible. For structural text lines, any axis is suitable, while for textual lines, we need to consider the horizontal axes. Given a set of detected lines $\mathcal{L}$, we estimate the homography matrix $\mathbf{H}$ by minimizing the following objective function:

$$\arg\min_{\mathbf{H}} \sum_{l \in \mathcal{L}} \Big[\|p(l, \mathbf{H})\|_2 \cdot \min_{r \in \mathcal{R}} \text{angle}(p(l, \mathbf{H}), r) \Big]^2 \quad (1)$$

with

$$p(l, \mathbf{H}) = \text{project}_{\mathbf{H}}(\text{approximate}(l)) \quad (2)$$

where approximate$(\cdot)$ maps a 2d line path the best fitting line using linear regression and $\mathcal{R}$ are the axis-aligned unit vectors to compare to. For text lines, $\mathcal{R}$ are the two horizontal unit vectors, while for structural lines, $\mathcal{R}$ are the four unit vectors. We apply the transformation to the background-removed image $\mathbf{W}_1$, which results in $\mathbf{W}_2$. See Figure 2 for an example.

3.4. Template-based Document Dewarping

We use the established template-based document dewarping model GeoTrTemplateLarge [12] to dewarp the image $\mathbf{W}_2$. It is based on an encoder-decoder architecture on top of deep embeddings of the input image and the template image. The attention mechanism allows for the model to connect template information with the unwarped image. Thereby, it integrates the template information implicitly. Our approach leverages the unwarping capabilities

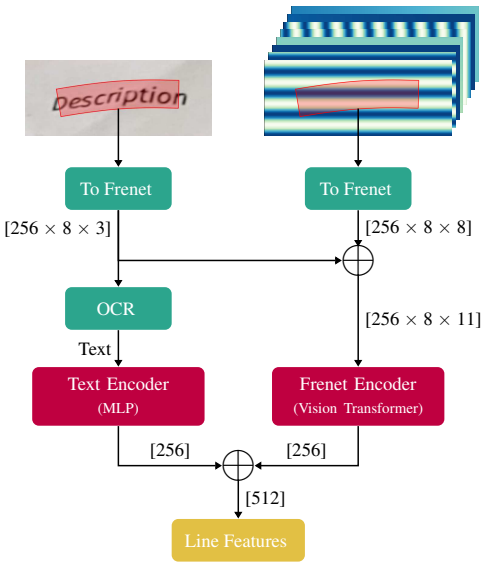


Figure 3. Visualization of the line encoding.

of GeoTrTemplateLarge before we add our more explicit line-based unwarping approach. We apply GeoTrTemplateLarge twice to the pre-unwarped image $\mathbf{W}_2$ to obtain the initial dewarped image $\mathbf{W}_3$.

3.5. Line Matcher

Given the initially dewarped image $\mathbf{W}_3$ and the template image $\mathbf{T}$, we aim to match the detected lines $\mathcal{L}(\mathbf{W}_3)$ to the template lines. To achieve this, we encode the line features and match them using the state-of-the-art local feature matcher LightGlue [23]. The line encoding consists of multiple parts: a visual encoding, a positional encoding, and an optional text encoding. See Figure 3 for a visualization of the line encoding. For the prior two encodings, we extend the input image $\mathbf{W}_3$ with a d-dimensional sinusoidal position encoding [31]. In our case, we set the dimensionality to 8, which results in an 11-dimensional input image. For each line $l \in \mathcal{L}(\mathbf{W}_3)$ and its width, we transform the line mask from Euclidean to Frenet coordinates [33]. The Frenet coordinates of a given pixel (x, y) are given by the distance d along the line and the distance o orthogonal to the line. When transforming the line mask to Frenet coordinates, we obtain a compact representation of the line $r_l \in \mathbb{R}^{256\times8\times11}$, which entails information about the line shape, position and visual appearance. We split the Frenet line encoding into 8 chunks of 8×8 pixels and apply a vision transformer [6] to the sequence of patches, which yields a deep line embedding with 256 dimensions. For text lines, we use a pretrained OCR model called docTR[1] to extract the text within the Frenet space and encode it with a simple MLP with two hidden layers, also resulting in a 256-dimensional vector. The full line representation is then given by the concatenation of the vision transformer and the text encoding.

For the matching of warped and template line embeddings, we use the local feature matcher LightGlue [23]. We train our model on the Inv3D dataset with the same hyperparameters as in the original paper. At inference time, we remove false matches by filtering out matches with a log assignment probability below the threshold t, matches without matching line types, and text matches without a common substring of at least three characters. We set the log assignment probability threshold to $t = -1$. This post-processing step reduces the number of false matches and thus improves the final dewarping result. We denote the matches as $\mathcal{M}(\mathbf{W}_3)$. See Figure 2 for an example.

3.6. Line Unwarping

Given the matches $\mathcal{M}(\mathbf{W}_3)$, we construct a dense transformation field to unwarp the image. A dense transformation field (also known as the forward map) is a 2D grid of vectors, where each vector represents the position shift of a pixel in the input image. We denote the forward map as $\mathbf{F} \in [0, 1]^{512\times512\times2}$. We start with the identity transformation and iteratively enhance the forward map by incorporating line correspondences. This is done by adding point correspondences and interpolating the missing vectors using Delaunay Triangulation [17]. The construction of the forward map $\mathbf{F}$ consists of three stages: known matches, support points, and unmatched lines. Figure 4 shows the construction of the forward map at each stage.

1. Project matches. In the first stage, we consider the known matches in descending order of assignment probability. For each match, we densely sample point correspondences along the warped and template line and add the points to the forward map, thus, forcing the pixels at the point correspondences to their location on the template. For text lines, we generate the point correspondences for both x and y coordinates. As for the structural lines, we only add either the x- or the y-coordinates to the forward map, depending on the template line direction. This is because structural lines only constrain the position in one direction. For example, if the structural line is horizontal, we only add the y-coordinates to the forward map.

2. Support points. For the second stage, we add support points outside the convex hull of all correspondences to reduce the interpolation error. Delaunay triangulation outside the convex hull of the correspondences (and within the image bounds) creates pixel shifts due to the triangular nature of the interpolation. Our goal with the support points is to retain the straightness of the image in those regions. To achieve this, we project the convex hull points to the left/right or top/bottom of the image space, depending on the gradient direction of each channel. Figure 4 shows an example of the support points.

[1] https://github.com/mindee/doctr

3. Unmatched lines. In stage three, we want to utilize the knowledge about all detected, but unmatched lines. To achieve this, we project the unmatched lines with the current forward map and straighten the lines in the unwarped space. This allows us to project the unmatched lines similar to stage 1 and thus, enforce the straightness of the unmatched lines in the final projection.

Since the line matcher occasionally produces false matches, we need to ensure that these false matches do not influence the final unwarping. To achieve this, we check the validity of the forward map during the iterative construction and reject all matches or lines that contradict our invariant. We define the invariant as follows: for each position in the forward map, the gradient must not exceed a range of valid values. We set the lower bound to 0, and the upper bound to 0.0025, respectively. This effectively limits the maximum compression and stretching for each part of the image and thus, disregards extreme cases. Due to the minor imprecisions of the line detector, our constructed forward map contains small artifacts that can lead to text distortion. In the final step, we apply a mean smoothing kernel to the backward map to retain the text readability.

We denote the final forward map as $\mathbf{F}$. By inverting the forward map, we obtain the backward map $\mathbf{B}$ and apply it to the pre-unwarped image $\mathbf{W}_3$. This yields the final dewarped image $\mathbf{U}$.

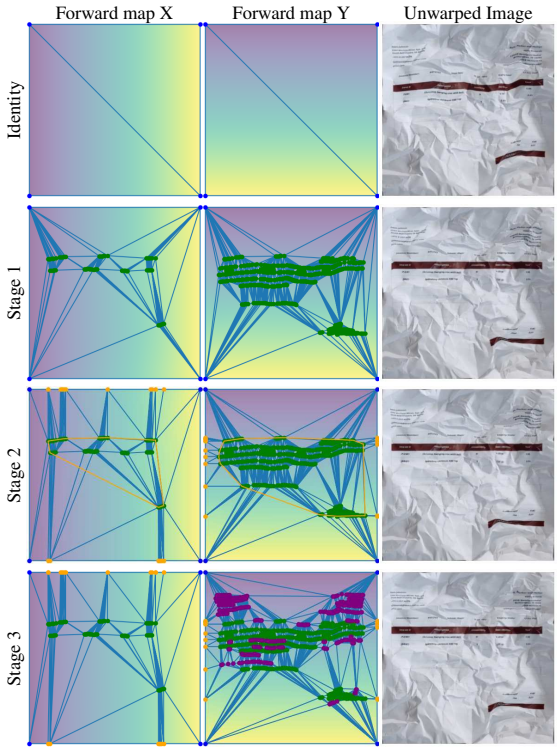


Figure 4. Construction of the forward map $\mathbf{F}$ for the line-based unwarping shown at multiple stages.

4. Evaluation

We evaluate our approach on the dataset Inv3DReal [12]. It consists of 360 invoice document images, subdivided into six deformation categories and three environment settings. We use a total of six different metrics to exhaustively evaluate the performance of our approach. Five of those metrics are established metrics for document image dewarping, while the sixth metric is a novel metric that we introduce in this work. The metrics can be divided into three categories: visual metrics, text-based metrics, and localized text-based metrics.

Visual metrics compare the unwarped image with the ground truth image, i. e. the invoice document before printing and photographing. Typically, there are three visual evaluation metrics: Multiscale structural similarity (MS-SSIM) [32], Learned Perceptual Image Patch Similarity (LPIPS) [40], and, Local Distortion (LD) [36].

The text-based metrics compare the extracted text from the unwarped image and the original image. For text extraction, we used the publicly available OCR engine docTR instead of the commonly used Tesseract OCR engine [29] due to its superior performance. The text-based metrics are Edit Distance (ED) [19] and Character Error Rate (CER).

Despite the merits of the previous metrics, they each have some limitations when it comes to evaluating the performance of document image dewarping. While the visual metrics are capable of identifying remaining geometric distortion, they do not capture the text readability at all and thus are not a suitable metric for the task of information extraction. The text-based metrics, on the other hand, focus on text readability, but they are overly sensitive to the positions of the recognized words. Structured documents, such as invoices, have text spread across the 2D document, and for the text-based metrics, these texts are linearized into a 1D sequence. Therefore, even slight changes in the position of the text can lead to a considerable change in the metric value. Additionally, text-based metrics do not capture the absolute positioning of the text in 2D space, which is crucial for the semantic interpretation of the text, especially if reference templates are available. To address these limitations, we introduce a novel metric that we call the *matched normalized Character Error Rate* (mnCER).

Given two images, the unwarped image I and the ground truth image G, we first extract the words from both images using the OCR engine docTR. Each word w is retrieved with its text t_w and its bounding box b_w in normalized image space $[0, 1] \times [0, 1]$. This yields two sets of words W_I and W_G for the unwarped and the ground truth image, respectively.

We then match the words from W_I to W_G based on the spatial locality of each word in W_G and each word in W_I.

We define the locality measurement as follows:

$$\text{locality}(b_i, b_g) = \begin{cases} \text{IoU}(b_i, b_g) & \text{if IoU}(b_i, b_g) > 0 \\ -d(b_i, b_g) & \text{if IoU}(b_i, b_g) = 0 \end{cases} \quad (3)$$

where b_i and b_g are the bounding boxes of the words w_i and w_g in W_I and W_G, respectively, and $d(b_i, b_g)$ is the minimal euclidean distance between the bounding boxes. The value range of the locality measurement is therefore $[-\sqrt{2}, 1]$ since the optimal intersection over union (IoU) is 1 and the largest distance possible is $\sqrt{2}$ within the normalized image domain. We then search a bipartite matching of words from W_I to W_G that maximizes the sum of the locality measurements for each word in W_I and each word in W_G using the Hungarian Assignment algorithm [15]. The assignment is denoted by $M \subseteq W_I \times W_G$. See Figure 5 for an example of the word matching.

Given the matched words, we then compute the normalized character error rate (nCER) as mentioned in [18] for each matched word w_i and w_g as follows:

$$\text{nCER}(w_i, w_g) = \frac{S + I + D}{S + I + D + C} \quad (4)$$

where S, I, D, and C are the number of substitutions, insertions, deletions, and correct characters, respectively. In contrast to the CER, the nCER ranges between 0 and 1, where 0 is the optimal score.

Our new metric mnCER is then defined as the average nCER overall matched words in M and a penalty value for each unmatched ground truth word in W_G:

$$\text{mnCER} = \frac{1}{|W_G|} \left[\sum_{(w_i, w_g) \in M} \text{nCER}(w_i, w_g) + |W_G| - |M| \right]$$

where $|W_G|$ is the number of words in W_G and $|M|$ is the number of matched words in M. The penalty value for unmatched words equals the worst possible nCER value, which is 1. The overall mnCER ranges between 0 and 1, where 0 is the optimal score.

The new metric mnCER solves the above-mentioned problems of the established metrics. It captures the text readability and the absolute positioning of the text in 2D space. Through the matching of words, it is robust to slight changes in the position of the text, and it is not overly sensitive to the positions of the recognized words.

5. Results

In this section, we present the results of our approach evaluated on the Inv3DReal dataset. We compare our approach to the identity mapping as a baseline, as well as the state-of-the-art document dewarping models GeoTrTemplateLarge [12], GeoTr [8] and the DewarpNet model [4] without refinement network.[2]

Figure 5. Example of the locality-based word matching for the mnCER calculation. Images (a) and (b) show the detected bounding boxes b_G and b_I, respectively. Image (c) displays the word matchings M.

Additionally, it is important to note that from our baselines, only the model GeoTrTemplateLarge utilizes a template image for dewarping. The other models are template-free and learn the dewarping based on the input image only. To the best of our knowledge, there is no other recent model that uses a template image for document image dewarping.

In the following, we present the quantitative results, followed by the qualitative results and an ablation study.

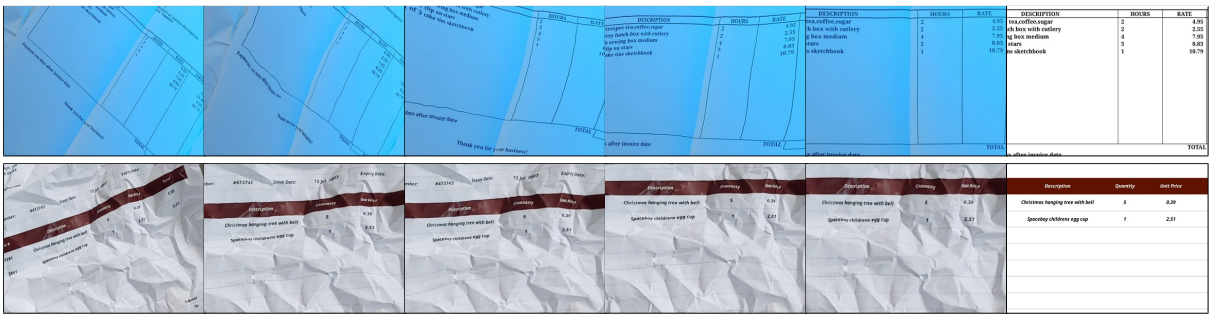
5.1. Quantitative Results

The quantitative results of our approach are shown in Figure and Table 1. When comparing our approach to the state-of-the-art models GeoTr, DewarpNet, and GeoTrTemplateLarge, and the baseline identity mapping, we can observe that our approach outperforms all models in all metrics. The comparison with the previous best model GeoTrTemplateLarge shows a significant improvement of 9.6 % in MS-SSIM, 11.8 % in LPIPS, 32.6 % in LD, 40.2 % in mnCER, 12.87 % in ED, and 11.68 % in CER. In particular, the metrics LD and mnCER show the largest improvements. Both metrics include the positioning of the elements in the unwarped image, indicating that the positioning of the text and visual elements improved significantly. The visual metrics MS-SSIM and LPIPS show a considerable improvement as well, indicating that the unwarped image looks closer to the original image. For the text-based met-

[2]Note that the latter two papers present an approach for document image dewarping in combination with illumination correction. Since our paper focuses on document image dewarping, we only compare against the dewarping part of the models.

	MS-SSIM ↑	LPIPS ↓	LD ↓	mnCER ↓	ED ↓	CER ↓
Identity	0.44 (0.10)	0.6 (0.10)	36.77 (13.54)	0.78 (0.14)	394.64 (140.81)	0.63 (0.19)
DewarpNet (w/o ref) [4]	0.55 (0.11)	0.42 (0.12)	25.27 (10.37)	0.58 (0.17)	254.92 (131.39)	0.41 (0.19)
GeoTr [8]	0.56 (0.11)	0.41 (0.12)	23.25 (10.16)	0.52 (0.18)	192.46 (118.88)	0.31 (0.18)
GeoTrTemplateLarge [12]	0.65 (0.12)	0.31 (0.13)	16.82 (10.31)	0.28 (0.18)	128.33 (99.09)	0.20 (0.15)
DocMatcher (ours)	**0.71 (0.12)**	**0.27 (0.12)**	**11.34 (8.26)**	**0.17 (0.13)**	**111.81 (79.56)**	**0.18 (0.13)**

Table 1. Quantitative results of our approach DocMatcher in comparison to the state of the art on the Inv3DReal dataset. All models were trained on Inv3D and evaluated on Inv3DReal. The values are given as mean and standard deviation.



(a) Input (b) DewarpNet (w/o ref) (c) GeoTr (c) GeoTrTemplateLarge (d) ours (e) Original

Figure 6. Qualitative evaluation of the state of the art and our approach based on selected samples of Inv3DReal.

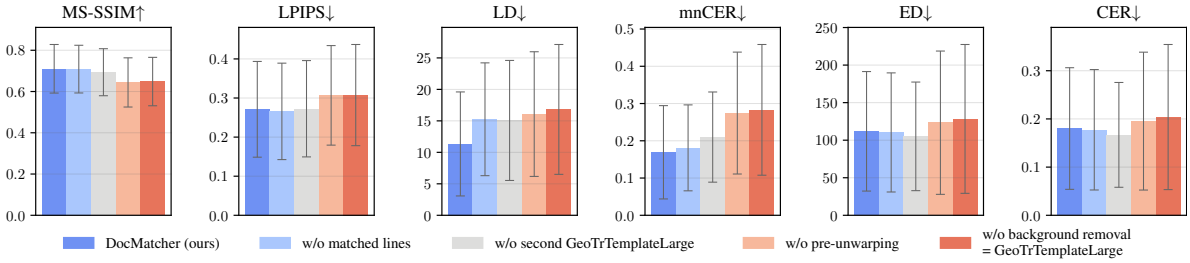


Figure 7. Ablation study for our proposed approach DocMatcher. We remove parts of the model piece by piece.

rics ED and CER, we also observe an improvement, but it should be noted that these metrics suffer from the linearization problem explained above and thus are not as reliable as the mnCER metric. It is also noteworthy, that the absolute numbers for the text-based metrics are significantly lower than the ones in the work of Hertlein *et al*. [12]. This is due to the change of the OCR engine from Tesseract to docTR, which handles challenging conditions better.

Overall, the results indicate that our approach is capable of dewarping document images more accurately than the previous state-of-the-art models.

5.2. Qualitative Results

We show the qualitative results of our approach compared to the state-of-the-art models for several selected samples in Figure 6. The first column displays the input image, whereas the last column presents the original image. The second to fifth columns show the results of the models DewarpNet (w/o ref) [4], GeoTr [8], GeoTrTemplateLarge [12], and our approach, respectively. We cropped the same region for all images to simplify the comparison. Thereby, it is easier to see the differences in the positioning of the text and visual elements.

In these examples, we can observe that our approach yields unwarped images, which are visually closer to the original image than the other models. The overall positioning of the elements and the straightness are improved, even under challenging lighting conditions (top row) or strong deformations (bottom row).

We want to point out, that the line-based unwarping stage can generate black artifacts at the border of the unwarped images. That effect can occur, when the line-based unwarping moves the pixels closer to the center and there are no pixels outside the image boundary that could fill the gap. We experimented with filling these artifacts by the color value of the nearest valid pixel. This variant yields a MS-SSIM of 0.72, a LPIPS of 0.26, a LD of 15.04, a mnCER of 0.17, an ED of 111.74 and a CER of 0.18. When comparing

it to our proposed variant, we observe a large change in LD, while the other metrics remain roughly the same. This sensitivity to the border artifacts indicates, that the metric LD is not robust to small changes in the image content.

5.3. Ablation Study

To evaluate the impact of the individual components of our approach, we conduct an ablation study. We ablate the full model piece by piece until we reach the base model GeoTrTemplateLarge [12] and evaluate the performance on the Inv3DReal dataset. We compare the full model to the following ablations: (1) without the line-based unwarping, (2) without the second inference of GeoTrTemplateLarge, (3) without the line-based pre-unwarping and finally without the background removal (4). The last ablation is equal to the model GeoTrTemplateLarge itself.

Figure 7 presents the results of the ablation. The first ablation - w/o line-based unwarping - shows significant degradation in LD and smaller degradation in mnCER. All other metric values are roughly the same as the full model. For the second ablation - w/o second GeoTrTemplateLarge - we find the largest degradation in mnCER. The third ablation - w/o line-based pre-unwarping - shows significant degradation in all metrics, especially in mnCER. And finally, the last ablation - w/o background removal - shows minor degradation in most metrics. While this step appears to have the least effect, it still contributes to the overall performance of the model. This indicates that all components are relevant to the full model. Considering the varying degrees of degradation among the metrics and ablations, it is reasonable to assume that each component has its own merits.

5.4. Susceptibility to Errors

In this section, we investigate the susceptibility of our approach to errors in intermediate stages of our pipeline. To that end, we design to experiments: The first experiment examines the influences of incorrect mask detection by the document detector. During inference, we rotate the detected document mask around its center by a random angle drawn uniformly from the interval $[-\alpha, \alpha]$. For the second experiment, we investigate the effect of incorrect line detections on the subsequent stages. Therefore, we randomly generate distractor lines during line detection by merging existing lines and add them to the detected lines.

The results of the error susceptibility experiment are shown in Table 2. The rotation of the document masks and the addition of distractor lines show little effect for small changes. For large changes, the impact of the added errors differs on the kind of error. Distractor lines seem to influence the overall performance hardly at all, while the mask rotation shows a decline in all metrics. However, it should be noted that even at the strongest rotation the visual metrics and mnCER are still better than the previous state of the art GeoTrTemplateLarge. The degradation in the text metrics might be caused by loss of information through incorrect masking. Overall, our approach shows robustness against errors in intermediate stages.

		MS-SSIM↑	LPIPS↓	LD↓	mnCER↓	ED↓	CER↓
Rotate	$\alpha = 25°$	0.68	0.30	12.0	0.23	145	0.23
	$\alpha = 20°$	0.68	0.29	11.6	0.21	141	0.23
	$\alpha = 15°$	0.69	0.28	11.6	0.18	129	0.21
	$\alpha = 10°$	0.70	0.28	**11.3**	**0.17**	124	0.20
	$\alpha = 5°$	**0.71**	**0.27**	**11.3**	**0.17**	116	0.19
	ours	**0.71**	**0.27**	**11.3**	**0.17**	112	**0.18**
Distractors	5%	**0.71**	**0.27**	11.5	**0.17**	115	**0.18**
	10%	0.70	0.28	11.5	**0.17**	113	**0.18**
	15%	0.70	0.28	11.5	0.18	**111**	**0.18**
	20%	0.70	0.28	11.8	0.18	112	**0.18**
	25%	0.70	0.28	11.7	0.19	117	0.19

Table 2. Results of the error susceptibility experiment.

6. Conclusion

In this work, we proposed a novel approach for document image dewarping called DocMatcher. It utilizes a template image to guide the dewarping process and achieves state-of-the-art results on the Inv3DReal dataset. We leverage the line orientation for a robust, content-based image pre-unwarping stage to reduce the pixel shift distance. Furthermore, we associate the lines between the warped and template image and leverage these associations to construct a dense transformation field that is capable of moving the document pixels to their correct position. In addition to that, we introduce a novel metric called Matched Normalized Character Error Rate (mnCER) that captures the text readability and the absolute positioning of the text in 2D space. Thereby, it removes the need for text linearization and thus, is more robust to slight changes in the position of the text.

We evaluate our approach on the Inv3DReal dataset and compare it to the state-of-the-art models. The results show that our approach outperforms the previous best model GeoTrTemplateLarge [12] in all metrics, especially in LD and mnCER. We conduct an ablation study to evaluate the impact of the individual components of our approach. Furthermore, we performed two experiments on the susceptibility of intermediate errors and showed that our approach is robust against these errors.

In the future, we will investigate the integration of more local features, such as points of interest to improve the correspondence-based unwarping.

References

[1] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, 8(6):679–698, 1986. 3

[2] Beiya Dai, Qunyi Xie, Yulin Li, Xiameng Qin, Chengquan Zhang, Kun Yao, Junyu Han, et al. Matadoc: Margin and text aware document dewarping for arbitrary boundary. *arXiv preprint arXiv:2307.12571*, 2023. 2

[3] Sagnik Das, Ke Ma, Zhixin Shu, and Dimitris Samaras. Learning an isometric surface parameterization for texture unwrapping. In *European Conference on Computer Vision*, pages 580–597. Springer, 2022. 2

[4] 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/CVF international conference on computer vision*, pages 131–140, 2019. 2, 6, 7

[5] Sagnik Das, Kunwar Yashraj Singh, Jon Wu, Erhan Bas, Vijay Mahadevan, Rahul Bhotika, and Dimitris Samaras. End-to-end piece-wise unwarping of document images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4268–4277, 2021. 2

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[7] Hao Feng, Shaokai Liu, Jiajun Deng, Wengang Zhou, and Houqiang Li. Deep unrestricted document image rectification. *IEEE Transactions on Multimedia*, 2023. 2

[8] Hao Feng, Yuechen Wang, Wengang Zhou, Jiajun Deng, and Houqiang Li. Doctr: Document image transformer for geometric unwarping and illumination correction. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 273–281, 2021. 2, 6, 7

[9] Hao Feng, Wengang Zhou, Jiajun Deng, Qi Tian, and Houqiang Li. Docscanner: Robust document image rectification with progressive learning. *arXiv preprint arXiv:2110.14968*, 2021. 2

[10] Hao Feng, Wengang Zhou, Jiajun Deng, Yuechen Wang, and Houqiang Li. Geometric representation learning for document image rectification. In *European Conference on Computer Vision*, pages 475–492. Springer, 2022. 2

[11] Felix Hertlein and Alexander Naumann. Template-guided illumination correction for document images with imperfect geometric reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop*, pages 904–913, 2023. 2

[12] Felix Hertlein, Alexander Naumann, and Patrick Philipp. Inv3d: a high-resolution 3d invoice dataset for template-guided single-image document unwarping. *International Journal on Document Analysis and Recognition (IJDAR)*, 26(3):175–186, 2023. 2, 3, 5, 6, 7, 8

[13] Xiangwei Jiang, Rujiao Long, Nan Xue, Zhibo Yang, Cong Yao, and Gui-Song Xia. Revisiting document image dewarping by grid regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4543–4552, 2022. 2

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 3

[15] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 6

[16] Jay Lal, Aditya Mitkari, Mahesh Bhosale, and David Doermann. Lineformer: Line chart data extraction using instance segmentation. In *International Conference on Document Analysis and Recognition*, pages 387–400. Springer, 2023. 3

[17] Der-Tsai Lee and Bruce J Schachter. Two algorithms for constructing a delaunay triangulation. *International Journal of Computer & Information Sciences*, 9(3):219–242, 1980. 4

[18] Kenneth Leung. Evaluate ocr output quality with character error rate (cer) and word error rate (wer), June 2021. Accessed on March 15, 2024. 6

[19] Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966. 5

[20] Heng Li, Xiangping Wu, Qingcai Chen, and Qianjin Xiang. Foreground and text-lines aware document image rectification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19574–19583, 2023. 2

[21] Pu Li, Weize Quan, Jianwei Guo, and Dong-Ming Yan. Layout-aware single-image document flattening. *ACM Transactions on Graphics*, 43(1):1–17, 2023. 2

[22] 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. 2

[23] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17627–17638, 2023. 4

[24] Shaokai Liu, Hao Feng, and Wengang Zhou. Rethinking supervision in document unwarping: A self-consistent flow-free approach. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 2

[25] 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

[26] Dong Luo and Pengbo Bo. Geometric rectification of creased document images based on isometric mapping. *arXiv preprint arXiv:2212.08365*, 2022. 2

[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

[28] Amir Markovitz, Inbal Lavi, Or Perel, Shai Mazor, and Roee Litman. Can you read me now? content aware rectification using angle supervision. In *Computer Vision–ECCV 2020:*

16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16, pages 208–223. Springer, 2020. 2

[29] 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. 5

[30] Floor Verhoeven, Tanguy Magne, and Olga Sorkine-Hornung. Uvdoc: Neural grid-based document unwarping. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 2

[31] Zelun Wang and Jyh-Charn Liu. Translating math formula images to latex sequences using deep neural networks with sequence-level training, 2019. 4

[32] Z. Wang, E.P. Simoncelli, and A.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 Vol.2, 2003. 5

[33] Moritz Werling, Julius Ziegler, Sören Kammel, and Sebastian Thrun. Optimal trajectory generation for dynamic street scenarios in a frenet frame. In *2010 IEEE international conference on robotics and automation*, pages 987–993. IEEE, 2010. 4

[34] Zhen Xu, Fei Yin, Peipei Yang, and Cheng-Lin Liu. Document image rectification in complex scene using stacked siamese networks. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1550–1556. IEEE, 2022. 2

[35] Chuhui Xue, Zichen Tian, Fangneng Zhan, Shijian Lu, and Song Bai. Fourier document restoration for robust document dewarping and recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4573–4582, 2022. 2

[36] 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. 5

[37] Fangchen Yu, Yina Xie, Lei Wu, Yafei Wen, Guozhi Wang, Shuai Ren, Xiaoxin Chen, Jianfeng Mao, and Wenye Li. Docreal: Robust document dewarping of real-life images via attention-enhanced control point prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 665–674, 2024. 2

[38] Jiaxin Zhang, Bangdong Chen, Hiuyi Cheng, Lianwen Jin, Fengjun Guo, and Kai Ding. Docaligner: Annotating real-world photographic document images by simply taking pictures. *arXiv preprint arXiv:2306.05749*, 2023. 2

[39] Jiaxin Zhang, Canjie Luo, Lianwen Jin, Fengjun Guo, and Kai Ding. Marior: Margin removal and iterative content rectification for document dewarping in the wild. *arXiv preprint arXiv:2207.11515*, 2022. 2

[40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[41] Weiguang Zhang, Qiufeng Wang, and Kaizhu Huang. Polardoc: One-stage document dewarping with multi-scope constraints under polar representation. *arXiv preprint arXiv:2312.07925*, 2023. 2