

DocAligner: Annotating Real-world Photographic Document Images by Simply Taking Pictures

Jiaxin Zhang¹, Bangdong Chen¹, Hiuyi Cheng¹, Fengjun Guo², Kai Ding², Lianwen Jin^{1*}

¹South China University of Technology, Guangzhou, China

²IntSig Information Co., Ltd., Shanghai, China

{eejxzheng, cbb5429}@gmail.com, eechenghiuyi@mail.scut.edu.cn, eelwjin@scut.edu.cn

{fengjun_guo, danny_ding}@intsig.net

Abstract

Recently, there has been a growing interest in research concerning document image analysis and recognition in photographic scenarios. However, the lack of labeled datasets for this emerging challenge poses a significant obstacle, as manual annotation can be time-consuming and impractical. To tackle this issue, we present DocAligner, a novel method that streamlines the manual annotation process to a simple step of taking pictures. DocAligner achieves this by establishing dense correspondence between photographic document images and their clean counterparts. It enables the automatic transfer of existing annotations in clean document images to photographic ones and helps to automatically acquire labels that are unavailable through manual labeling. Considering the distinctive characteristics of document images, DocAligner incorporates several innovative features. First, we propose a non-rigid pre-alignment technique based on the document's edges, which effectively eliminates interference caused by significant global shifts and repetitive patterns present in document images. Second, to handle large shifts and ensure high accuracy, we introduce a hierarchical aligning approach that combines global and local correlation layers. Furthermore, considering the importance of fine-grained elements in document images, we present a details recurrent refinement module to enhance the output in a high-resolution space. To train DocAligner, we construct a synthetic dataset and introduce a self-supervised learning approach to enhance its robustness for real-world data. Through extensive experiments, we demonstrate the effectiveness of DocAligner and the acquired dataset. Datasets and codes will be publicly available.

1. Introduction

In recent years, researchers have made significant strides in document image analysis and recognition. While previ-

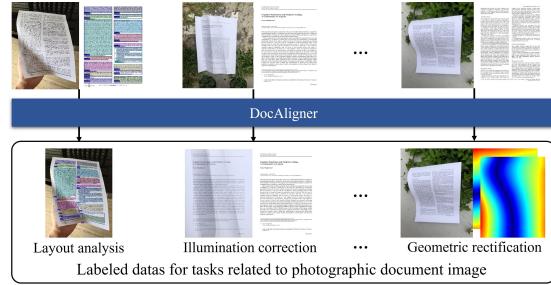


Figure 1. Automatic annotation for real-world photographic document images via DocAligner. All you need to do is take pictures.

ous studies predominantly focused on clean document images obtained from digital-born sources or flat-bed scanners [20, 21, 29, 38, 47], there is a growing interest among researchers in addressing the challenges posed by more realistic photographic scenarios [23, 25, 27, 37]. However, progress in this field has been hindered by the limited availability of labeled photographic data. This data scarcity can be attributed to several reasons. Firstly, automatic labeling methods [3, 20, 21, 47] designed for clean document images are not suitable for photographic scenarios, necessitating costly and time-consuming manual labeling. Secondly, certain tasks such as illumination correction and geometric rectification are extremely challenging to annotate manually.

To address the aforementioned issues, we propose DocAligner, a novel method that significantly simplifies manual annotation to just taking pictures. DocAligner achieves this by establishing dense correspondence between photographic document images and their clean counterparts, which is a new perspective in the context of document artificial intelligence. As shown in Fig. 1, for tasks that already have a large amount of labeled clean images (such as layout analysis, table detection, and table recognition), annotations can be transferred to corresponding photographic images. For tasks that cannot rely on existing labeled data and re-

quire extensive labeling efforts, we can automatically generate labels following the dense correspondence. In other words, to annotate photographic data, it is only necessary to print a clean document image and take a picture. DocAligner can perform the remaining tasks automatically.

While dense correspondence has been extensively explored in the realm of natural images, utilizing pre-existing models designed for natural images in the context of document analysis faces performance degradation issues resulting from the distribution gap. To address this issue, we make novel designs for DocAligner to achieve dense correspondence exclusively for document images. Pre-alignment becomes necessary for document pairs that exhibit significant global misalignment and contain repetitive patterns. Unlike rigid transformations such as affine and homography commonly used for pre-alignment in natural images, these methods are not suitable for document pairs with non-rigid deformation. Therefore, we propose the use of a thin plate spline (TPS) [2] non-rigid transformation, inspired by the advancements in geometric rectification techniques [24, 45]. To handle significant shifts and ensure high accuracy, we adopt hierarchical alignment that combines global-local correlation and coarse-to-fine flow prediction. Furthermore, document images possess a more intricate structure, with the details at the character level being crucial. Consequently, obtaining higher-resolution output flows becomes necessary in DocAligner. To address this, we propose a details recurrent refinement module that operates in the detail-rich and high-resolution space. To mitigate memory consumption in high-resolution processing, we conduct refinement recurrently using the memory-efficient ConvGRU [32].

To train DocAligner, we develop a synthetic dataset comprising triplets of photographic document images, clean document images, and flow fields. The clean document images are derived from PDF files. We warp these clean images with randomly-generated flows and then deteriorate them using collected shading maps to synthesize photographic images. Additionally, to further improve DocAligner’s performance on real data, we propose a self-supervised learning approach. Experimental results demonstrate the superiority of DocAligner compared to existing methods. Furthermore, we assess the effectiveness of the acquired dataset in multiple tasks related to photographic document images, including layout analysis, illumination correction, and geometric rectification.

In summary, our contributions are as follows:

- For the first time, we explore the dense correspondence task in the context of document artificial intelligence, by which we ease the data dilemma countered by tasks related to photographic document images.
- We propose DocAligner for document image dense

correspondence, in which we design non-rigid pre-alignment, hierarchical alignment with global and local correlation, and details recurrent refinement. We also develop a synthetic dataset and a self-supervised learning approach that is easy to implement and helps to improve generalization.

- DocAligner achieves superior performance compared to existing methods. Additionally, we validate the effectiveness of the dataset we acquired for related tasks.

2. Related works

2.1. Dense correspondence

Dense correspondence of paired images has been extensively studied for natural images in recent years [11, 14, 18, 26, 30, 33, 34, 36]. Given an image pair (I_s, I_t) with a size of $H \times W$, dense correspondence aims to predict a flow field $f \in \mathbb{R}^{H \times W \times 2}$, which relates the source I_s to the target I_t . According to differences within the paired images, dense correspondence can be categorized into optical flow [14, 33, 34], geometric correspondence [26, 30, 36] and semantic correspondence [11, 18, 30, 36]. Document image pairs with large displacements and significant appearance transformations are more relevant to geometric correspondence, where pairs usually exhibit different views of the same scene or are captured by different cameras on different occasions.

Melekhov et al. [26] proposed DGC-Net, a neural network with a global correlation layer that can handle large displacement. However, due to the large memory footprint of this layer, the input image resolution for DGC-Net is constrained to 240×240 . Such a coarse resolution is insufficient for representing a document with fine-grained content. Glu-Net [36] takes a more elegant approach by performing global correlation in coarse resolution and local correlation in fine resolution, resulting in better performance on high-resolution input. Truong et al. proposed GOCor [35], a new optimizable correlation layer, to enhance the correlation robustness of similar and low-textured regions. Some self-supervised methods are also introduced to make models more robust on real-world data. RANSAC-Flow [31] adopts a two-stage framework where coarse alignment based on RANSAC [10] is followed by a fine alignment based on a deep model trained with self-supervision. However, it is sophisticated and hard to implement due to its multi-task optimization for cycle consistency, matchability, and reconstruction. DMP [13] optimizes the untrained matching networks on a single pair of images. But it is less practical to focus solely on an input pair. Although achieving promising results on natural images, the above-mentioned methods remain sub-optimal in the context of document images.

2.2. Document analysis and recognition in photographic scenarios

Document layout analysis (DLA). DLA aims to identify the regions of interest in an unstructured document and determine the role of each region. Previous studies have primarily focused on digital-born document images that are relatively easy to label by parsing PDFs and analyzing the corresponding source codes [1, 17, 42, 46] such as LaTeX and XML. However, more and more photographic document images are emerging, which the existing automatic labeling methods cannot cope with. Consequently, while millions of labeled clean document images are available, there are limited datasets for photographic scenarios. Although it is possible to annotate manually (by annotating bounding boxes and classes like title, author, list, abstract, paragraph, table, figure, etc.), it is expensive, especially considering geometric deformation and the large number of objects in photographic document images.

Illumination correction. Illumination correction seeks to eliminate degradation caused by uncontrolled illumination, enhancing readability and facilitating following optical character recognition (OCR) engines [6, 8, 22]. Nevertheless, obtaining labels for this task, i.e., the illumination-corrected image, is challenging due to the dense annotation density. An alternative approach to obtain labeled data is to capture the document under different illuminations while keeping its relative position fixed with the camera. However, source documents are not always available, and the variety and scale of datasets obtained in this way are minimal. Thus, most recent learning-based illumination correction methods [5, 6, 8, 22] can only be trained on synthetic data, whose realism and diversity remain unsatisfactory. Further discussions are included in Section 4.4.

Geometric rectification. This task aims to flatten document images that suffer from curves, folds, crumples, etc. A dewarping map is required to sample from distorted input to obtain the rectified result. This dewarping map indicates the correspondence between pixels in the desired rectified results and the distorted inputs. However, such a dewarping map is an extremely dense annotation that can be almost impossible to obtain manually [40]. Consequently, many learning-based methods have to resort to synthetic data. Failure to obtain the dewarping map annotation means that real-world data can only be used for weak supervision [24, 41].

3. Methodology

As shown in Fig. 2, the proposed DocAligner seeks to correlate the photographic document image I_s with its clean counterpart I_t . To achieve this, a pre-alignment module is first utilized to obtain I'_s , which is accomplished using an edge-based non-rigid transformation. Pre-aligned pairs

(I'_s, I_t) are then fed into a shared feature extraction backbone to extract multi-scale features, which are then used to predict flows hierarchically. Finally, a refinement module is designed to refine flow details recurrently in high-resolution space and output a flow with the same size as the input pairs.

3.1. Non-rigid pre-alignment

Photographic document images often suffer from significant global misalignment caused by the varying camera angles and paper deformations. Such misalignment combined with repetitive patterns in a document impede precise correlations. Pre-aligning the images before carrying out fine-grained correlation can resolve this issue. Traditional affine and homography rigid transformations for natural images [15, 30, 31] are not suitable for document pairs that are with both rigid and non-rigid deformations. In this paper, we focus on the advancements in document image rectification which utilize the document edge information [45]. Specifically, as illustrated in Fig. 2, we first extract the edge of the document through semantic segmentation. We then detect the four corners and equidistant points on four edges based on edge information. We map these detected points to their pre-defined reference counterparts in a quadrilateral. Using these paired points, we apply the TPS non-rigid transformation to obtain pre-aligned I'_s .

3.2. Hierarchical alignment

As shown in Fig. 2, we feed the pre-aligned pairs (I'_s, I_t) into a shared feature extraction backbone which produce multi-scale features $X^s = \{X_1^s, X_2^s, X_3^s, X_4^s\}$ and $X^t = \{X_1^t, X_2^t, X_3^t, X_4^t\}$ for I'_s and I_t , respectively. In the hierarchical alignment module, we predict and fine-tune the flow from low to high resolution, i.e., from level $l = 1$ to $l = 3$. The flow f_l between a pair of features (X_l^s, X_l^t) at level l is calculated by

$$f_l = \text{up}(f_{l-1}) + \text{decoder}_l(C_l, \text{up}(f_{l-1})), \quad (1)$$

where $\text{up}()$ is a bilinear up-sampling function and $\text{decoder}_l()$ is a lightweight fully convolution neural network. The detailed architecture of $\text{decoder}_l()$ can be found in the supplementary materials. Furthermore, C_l refers to the correlation map obtained through global or local correlation layer $\mathbf{C}_{\mathbf{G/L}}$:

$$C_l = \mathbf{C}_{\mathbf{G/L}}(\tilde{X}_l^s, X_l^t). \quad (2)$$

Here \tilde{X}_l^s is obtained by warping X_l^s toward X_l^t :

$$\tilde{X}_l^s(x) = X_l^s(\mathbf{x} + \text{up}(f_{l-1})(\mathbf{x})), \quad (3)$$

where \mathbf{x} denotes the image coordinate. Additionally, the initial flow f_0 is a zero-filled map.

The correlation layer, also known as cost volume, is essential in current state-of-the-art dense correspondence methods, as it represents the similarities between spatial elements in both reference and query features. The global

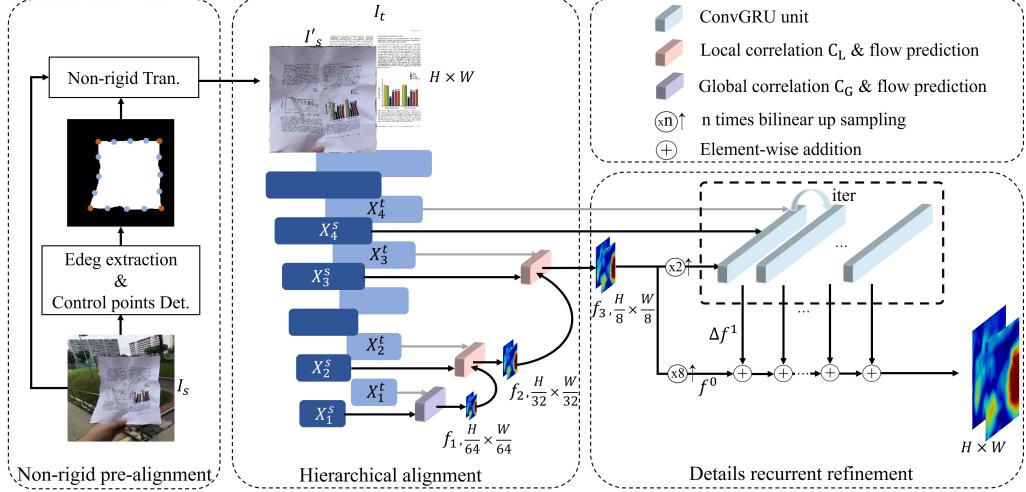


Figure 2. Overall architecture of DocAligner. Considering photographic I_s and clean I_t , DocAligner aims to densely correlate each pixel in I_t with I_s .

correlation layer calculates the scalar product between each feature vector in the reference features $X^r \in \mathbb{R}^{H^r \times W^r \times D}$ and all the vectors in the query features $X^q \in \mathbb{R}^{H^q \times W^q \times D}$, as following:

$$\mathbf{C}_G(X^r, X^q)_{ij} = (x_i^r)^T x_j^q, \quad (4)$$

where $x_i^r \in \mathbb{R}^D$ and $x_j^q \in \mathbb{R}^D$ are the i -th and j -th vector in X^r and X^q , respectively. This layer, denoted as $\mathbf{C}_G \in \mathbb{R}^{H^r W^r \times H^q W^q}$, represents similarities between all locations in the reference and query features and can handle large displacements. However, its computational complexity and memory consumption increase quadratically with feature size, rendering it only suitable for low-resolution features. In contrast, the local correlation layer computes the scalar product between vectors within a constrained distance:

$$\mathbf{C}_L(X^r, X^q)_{id} = (x_i^r)^T x_{i+d}^q, \quad \|d\| \leq R \quad (5)$$

where R is the pre-defined constant ratio. $\mathbf{C}_L \in \mathbb{R}^{H^r W^r \times (2R+1)}$ is more computationally efficient but unsuitable for correlating feature pairs with large displacements. In this paper, we apply the global correlation layer to the lowest resolution features (i.e., $l = 1$) and the local correlation layer with $R = 9$ to the remaining levels, enabling hierarchical alignment that can eliminate large global displacements in low resolution and focus precise correlation in high resolution.

3.3. Details recurrent refinement

The final output flow field of natural images is usual 1/4 the size of the input image. A flow field with a larger size will not bring much more improvement [4, 7, 33, 36]. For some natural scenarios, low-resolution input is enough to

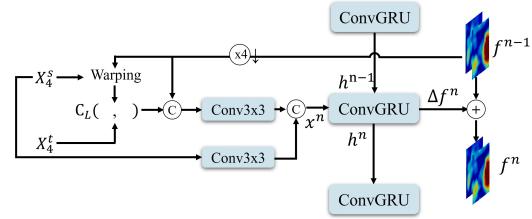


Figure 3. The details recurrent refinement module.

obtain good performance [16, 26, 31]. Nevertheless, this is not the case for document images because of their fine-grain elements. To address this issue, we propose a details refinement module to obtain the output flow field with the same size as input pairs. As shown in Figs. 2 and 3, this refinement module refines the output in the highest resolution space and employs a recurrent ConvGRU unit [32] to reduce memory usage. At each time step n , the input of the GRU unit is x^n , which is obtained by

$$x^n = [\text{Conv3x3}(X_4^s), \text{motion}], \quad (6)$$

$$\text{motion} = \text{Conv3x3}([\text{down}(f^{n-1}), \mathbf{C}_L(\tilde{X}_4^s, X_4^t)]), \quad (7)$$

$$\tilde{X}_4^s(x) = X_4^s(x + \text{down}(f^{n-1})(x)). \quad (8)$$

$\text{Conv3x3}()$ refers to a 3×3 convolutional layer, $\text{down}()$ is a down-sampling function, and $\tilde{X}_4^s(x)$ is obtained by warping X_4^s toward X_4^t .

The GRU unit updates the hidden state from the input x^n and the previous hidden state h^{n-1} , similar to a method in [32]. The updated hidden state h^n is used to predict the residual flow $\Delta f^n \downarrow$ using a two-layer convolutional neural network.

However, the size of $\Delta f^n \downarrow (\frac{H}{4} \times \frac{W}{4})$ is smaller than that of the input image, so we introduce a learnable upsampling

method. We predict 16 weight matrices of size 3×3 for each pixel in $\Delta f^n \downarrow$, thus obtaining the weight map with size of $\frac{H}{4} \times \frac{W}{4} \times 16 \times 3 \times 3$, which can be reshaped as $\frac{H}{4} \times \frac{W}{4} \times 144$. We obtain it by feeding $\Delta f^n \downarrow$ and h^n to another two-layer convolutional neural network (refer to supplementary materials for more details). We then obtain each upsampled pixel using a weighted sum over the 3×3 neighborhood in $\Delta f^n \downarrow$ and finally obtain the desired upsampled residual flow Δf^n . The updated flow can be obtained by

$$f^n = f^{n-1} + \Delta f^n. \quad (9)$$

Bilinearly up-sampled flow f_3 from the hierarchical alignment module is set as the initial flow f^0 for Eq. 8 and 9. Initial hidden state h^0 is transformed from X_4^s through a 1×1 convolution layer. We set the iterations for refinement as 7.

3.4. Self-supervision learning with real data

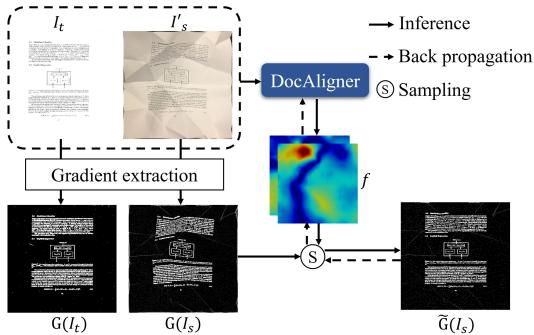


Figure 4. Scheme of self-supervision on real data. The dotted lines indicate the back propagation of gradient.

We train DocAligner on synthetic data as described in Section 4.1, but we observe a distribution gap between real-world and synthetic data. To tackle this problem, we propose a self-supervised approach to improve DocAligner’s robustness, as depicted in Fig. 4. This approach involves providing DocAligner with paired clean document image I_t and pre-aligned photographic image I'_s from the real world, and is formulated as

$$\hat{\theta} = \operatorname{argmin}_{\theta} \left(\left\| \tilde{\mathbf{G}}(I'_s) - \mathbf{G}(I_t) \right\|_1 \right), \quad (10)$$

$$\tilde{\mathbf{G}}(I'_s) = \mathbf{G}(I'_s)(\mathbf{x} + f(\mathbf{x})), \quad (11)$$

$$f = \text{DocAligner}(I_t, I'_s; \theta). \quad (12)$$

Here θ is the parameters of DocAligner, f is the predicted flow field, \mathbf{x} is the image coordinates, and \mathbf{G} is the Sobel operator to extract gradients. It is hard to directly solve Eq. 10, so we approximate it using a gradient descent algorithm. Considering the inefficiency of individual optimization of each sample, similar to prior art [31], we optimize our network on the entire test set before testing and then perform inference using the optimized network. It should

be noted that our self-supervision process only involves the input data and does not include any ground-truth flow field for supervision.

4. Experiments

4.1. Dataset

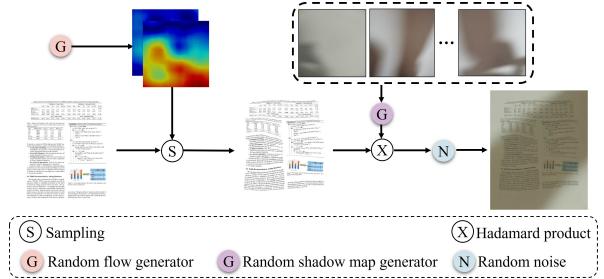


Figure 5. The pipeline for synthesizing DocAlign12K.

Due to the lack of publicly available training data, we develop a synthetic dataset named DocAlign12K, the synthetic pipeline of which is shown in Fig. 5. Firstly, we collect the PDF files from the Internet and then convert them into clean document images. Next, we randomly generate flow fields for each of these images. More detailed procedures and parameter settings are available in the supplementary materials. Using the generated flow fields, we warp clean images to obtain geometrically distorted images. Based on the Lambertian assumption, we consider an image I as a composition of reflectance R and shading S , i.e., $I = R \otimes S$, where \otimes denotes the Hadamard product. We collect 500 real shadings by capturing backgrounds without texture under various illumination conditions. We select one of these shadings randomly and perform random cropping, rotation, and color shifts to obtain S . Finally, we treat the geometrically distorted image as R to obtain the final image I . Additionally, we apply random JPEG compression, Gaussian noise, blur, etc., to better simulate the degradations. We obtain 12K samples (with triplets of clean documents, photographic documents, and flow fields) and split them into training (10K) and testing (2K) sets.

4.2. Implementation details

We implement DocAligner in the PyTorch framework [28], and train it on two NVIDIA 2080Ti GPUs with a batch size of 4. The widely-used Adam optimizer [19] is adopted. The initial learning rate is set to 1×10^{-4} and reduced by a factor of 0.3 after every 30 epochs. We train each model for 100 epochs. The shared feature extraction backbone is a ResNet-18 pre-trained on ImageNet. The input size is set to 1024×1024 . When trained with DocAlign12K, flows from all hierarchical levels and refinement iterations are supervised by the ground-truth flow with $L1$ loss.

4.3. Comparison with state-of-the-art

We make comparisons with DGC-Net [26], Glu-Net [36], and Glu-GOCor [35]. All these methods are trained with DocAlign12K. To ensure a fair comparison, we apply our non-rigid pre-alignment method to these methods since DocAlign12K does not consider the background margin. Note that the input spatial resolution for DGC-Net is set to 240×240 , while that for Glu-Net and Glu-GOCor is set to 1024×1024 .

We first evaluate DocAligner and the above-mentioned methods on the testing set of DocAlign12K. Similar to dense correspondence in natural images [26, 35, 36], we use average endpoint error (AEPE) and percentage of correct keypoints (PCK) [43] as metrics. AEPE measures the averaged Euclidean distance between predicted and ground-truth flow fields over all pixels. PCK is defined as the percentage of the correctly estimated points that are within a certain Euclidean distance threshold (in pixels). Results in Table 1 demonstrate that DocAligner outperforms previous methods in all metrics by a considerable margin. Particularly, DocAligner obtains a relative improvement of 23.4% over the second-best method in PCK-1px, indicating DocAligner’s ability to achieve precise correlation.

Table 1. Comparisons on DocAlign12K’s testing set.

	AEPE \downarrow	PCK-1px (%) \uparrow	PCK-5px (%) \uparrow
DGC-Net [26]	47.39	6.98	15.67
Glu-Net [36]	1.82	51.04	93.74
Glu-GOCor [35]	1.54	62.15	94.49
DocAligner	1.09	76.63	96.36

We then use DocUNet [25] benchmark to further validate DocAligner’ performance, which consists of clean scanned and geometrically-distorted photographic document images. We use the predicted flow field to warp the pre-aligned photographic image: $\tilde{I}'_s = I'_s(\mathbf{x} + f(\mathbf{x}))$, and then assess the alignment between the target clean scanned image I_t and \tilde{I}'_s . The better the alignment, the better the model’s performance. Multi-scale structural similarity (MS-SSIM) [39] and align distortion (AD) [24] are adopted as evaluation metrics. MS-SSIM is a widely used metric that measure perception-based similarity between two images. AD aims to measure the local distortion between two images via dense SIFT flow, which is improved from local distortion (LD) [44] by excluding the noise in low-textured regions and excluding the effect of subtle global transformations. Quantitative and qualitative results are given in Table 2 and Fig. 6, respectively. Although DGC-Net can achieve seemingly feasible global alignment, it fails to warp the character details because of the detail-lacking input. The correlation layers in Glu-GOCor are more robust toward repetitive patterns, so it exhibits improvement when compared to Glu-Net. DocAligner achieves superior per-

Table 2. Comparisons with state-of-the-art geometric correspondence methods on DocUNet dataset. $SSFT^{10}$ denotes Self-Supervised Fine-Tuning on the entire test set with 10 epochs before testing.

Methods	MS-SSIM \uparrow	AD \downarrow	Parameters (M)	Run-time (s)
DGC-Net [26]	0.6177	0.3137	68.47	0.48
Glu-Net [36]	0.7728	0.1186	94.17	0.75
Glu-GOCor [35]	0.7862	0.0938	94.17	0.85
DocAligner	0.8058	0.0486	103.8	0.93
DocAligner $SSFT^{10}$	0.8232	0.0445	103.8	0.93

Table 3. Performance on DocUNet dataset when solely trained with self-supervision.

Methods	MS-SSIM \uparrow	AD \downarrow
RANSAC-Flow [26]	0.6746	0.4700
DocAligner $SSFT^{10}$	0.7864	0.0918

formance when compared to the above-mentioned methods. Moreover, the gains are further boosted when our self-supervised method is applied.

In order to validate the effectiveness of our self-supervised approach, we compare it with RANSAC-Flow [31], which is another self-supervised method. Similar to the settings in [31], we do not train models on our DocAlign12K but solely adopt the self-supervised training on real data. We train RANSAC-Flow and DocAligner on the WarpDoc [41] dataset, which consists of 1020 pairs of photographic and clean document images, and then fine-tune and test them on DocUNet. Results shown in Table 3 demonstrate the superiority of DocAligner. We observe that the poor performance of RANSAC-Flow partially contributes to its rigid pre-alignment, which can not obtain satisfying pre-alignment results. Besides, our self-supervised method is more easy to be implemened compared to sophisticated hierarchical learning for multi-task optimization in RANSAC-Flow.

4.4. Applications of DocAligner

To further validate the feasibility and application value of our approach, we annotate photographic document images for document layout analysis (DLA), illumination correction, and geometric rectification using our DocAligner $SSFT^{10}$, and validate the effectiveness of the acquired data. For brevity, we will omit the subscript $SSFT^{10}$ in the following discussion.

Document layout analysis. We use DocAligner to transfer annotations from an existing dataset for document layout analysis (DLA) to photographic images. To accomplish this, we randomly select 2200 samples from PubLayNet [47], a large-scale DLA dataset of clean document images. We then print these images and capture them in various environments before using DocAligner to correlate photographic and clean pairs. The resulting flow from Do-

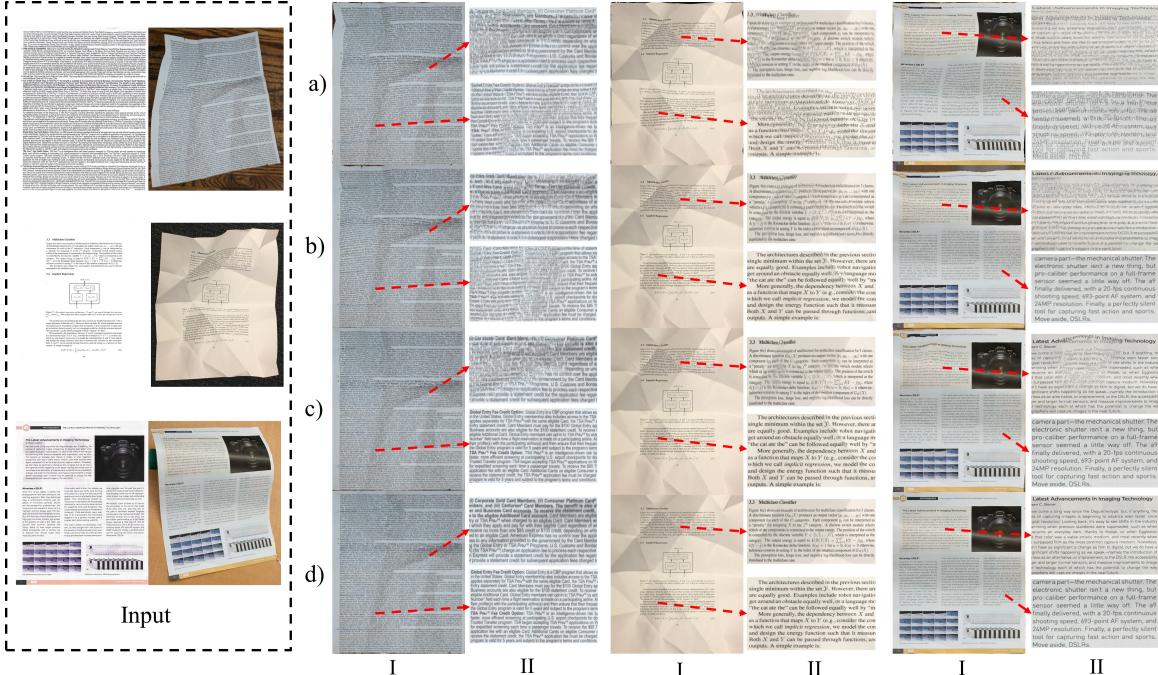


Figure 6. Input pairs are shown on the left. Warped results are shown in I. In II, we overlap targets on warped results and show local details. Top to bottom show results from a) DGC-Net [26], b) Glu-Net [36], c) Glu-GOCor [35], and d) DocAligner^{SSFT10}.

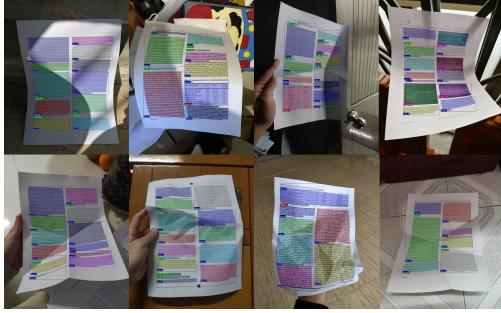


Figure 7. Some samples from our acquired DLA dataset.

Table 4. Performance on photographic testing set when trained with different data. Clean images are original from PubLayNet. S and G represent shadow and geometric synthesis, respectively, the same as DocAlign12K.

Training data	Type	Num.	mAP(@0.5-0.95)
Clean images	-	2K	8.0
Clean images + S	Synthetic	2K	36.9
Clean images + G	Synthetic	2K	21.5
Clean images + G + S	Synthetic	2K	49.7
Clean images + G + S	Synthetic	20K	61.9
Data from DocAligner	Real	2K	68.0

cAligner enabled us to transform the coordinates of bounding boxes and masks to their photographic counterparts, allowing us to obtain the annotations we needed. It is worth

Table 5. Performance on DocUNet dataset when trained with different dataset. Results in the first line represent images without illumination correction (i.e., input images for model).

Network	Training data	Num.	SSIM↑	PSNR↑
-	-	-	0.7065	12.90
illNet	DocProj	2450	0.7139	15.74
	Dataset from DocAligner	800	0.7504	16.78

noting that traditional manual labeling processes typically cost 5-15 minutes per image, but our approach reduces this to approximately 0.15 minutes per image. Some acquired samples are shown in Fig. 7, which demonstrates DocAligner’s ability to generate diverse data with high-quality annotations. To validate the effectiveness of the acquired data, we use it to train a Mask R-CNN [12]. We randomly select 200 samples, while the remaining samples form the training set. We manually inspect the testing set and make adjustments to any incorrect annotations to ensure labeling accuracy. The results, shown in Table 4, illustrate the significant superiority of our acquired training data compared to synthetic data. Additional visualization results can be found in the supplementary materials. Furthermore, it is possible to label other detection tasks involving bounding box annotations, such as table and text line detection, using a similar approach.

Illumination correction. We utilize our DocAligner to

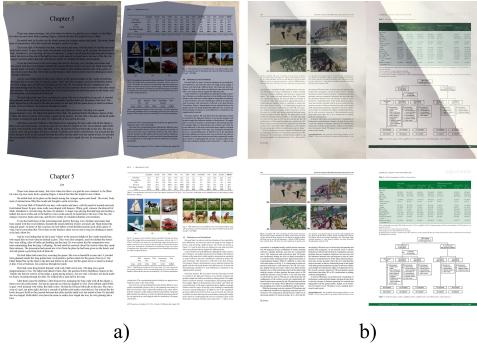


Figure 8. a) Examples from synthetic DocProj dataset and b) examples from our acquired real-world dataset.

Table 6. Performance when trained with different data. R and S denote real and synthetic, respectively.

Model	Traning data	Type	Num.	MS-SSIM↑	AD↓
DocUNet [25]	Ma et al. [25]	S	100K	0.4157	0.4957
DocProj [22]	Li et al. [22]	S	1K	0.2531	0.9278
DDCP [40]	Xie et al. [40]	S	30K	0.4189	0.5071
DewarpNet [5]	Doc3D [5]	S	100K	0.4057	0.5187
DocTr [8]	Doc3D [5]	S	100K	0.4649	0.4708
PaperEdge [24]	DIW [24]+Doc3D [5]	R+S	2.3K+100K	0.4523	0.3901
Transformer-based	Dataset from DocAligner	R	2.5K	0.4897	<u>0.4226</u>

annotate the WarpDoc [41] dataset, which consists of clean scanned document images and geometrically-distorted photographic document images. However, we exclude the 'Incomplete' subset since our approach does not currently support this document type, as explained in the supplementary material. Furthermore, we exclude images with excessively large rotation angles. It is worth noting that such problematic images can be avoided during the photography process by informing the collectors in advance. Using the obtained flow field, we warp the photographic source images towards their clean target images to generate paired data for illumination correction. In total, we obtain 800 pairs. Fig. 8 presents examples of our acquired data alongside synthetic data from DocProj [22]. It can be observed that the synthetic data lacks diversity and realism. For quantitative comparisons, we train illNet [22] with our acquired data and DocProj data respectively. The resulting model corrects the illumination of geometrically-rectified DocUNet images from our DocAligner^{SSFT10} (i.e., results depicted in the last row of Table 2). Table 5 presents the SSIM and PSNR values between the results obtained by illNet and the corresponding scanned clean images. The results indicate that while model trained with synthetic data yield improvement, it still lag behind model trained with real data, even if the real dataset is smaller in size. Supplementary materials include qualitative comparisons of models trained with different datasets.

Geometric rectification. We combine image pairs in

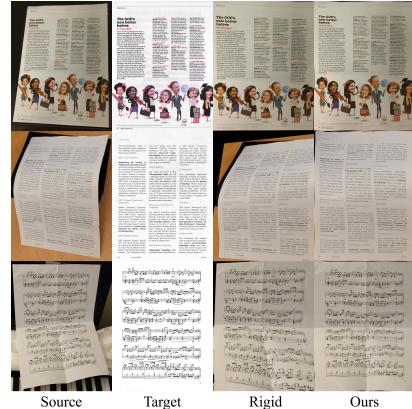


Figure 9. Comparisons between results from rigid pre-alignment and our non-rigid pre-alignment.

Table 7. Ablation studies on our non-rigid pre-alignment and details recurrent refinement module.

Non-rigid pred-alignment	Details recurrent refinement	MS-SSIM↑	AD↓
✗	✓	0.7578	0.1099
✓	✗	0.7910	0.0756
✓	✓	0.8058	0.0486

DIB [9], WarpDoc [41], DocUNet [25], and our collected DLA data together, from which we randomly select 300 pairs as a testing set. The rest are correlated using DocAligner to get flow fields as annotations. We finally get a training set with 2.5K samples. Inspired by the success achieved by DocTr [8], we adopt the Transformer as our dewarping network without extra sophisticated designs. We train it using our constructed training set. As demonstrated in Table 6, model trained using our acquired data yield promising results. Our model, based on the vanilla Transformer architecture, surpasses previous meticulously-designed approaches that were trained on large-scale synthetic data. This outcome confirms the effectiveness of the real data we acquired.

4.5. Ablation studies

Non-rigid pre-alignment. We replace non-rigid pre-alignment module with homography transformation and evaluate on DocUNet. We use deep features from a pre-trained ResNet-50 to represent two paired images and obtain sparse correspondences based on cosine similarity. Then the RANSAC algorithm [10] is applied to fit a homography. Results in Table 7 indicate that replacing non-rigid pre-alignment leads to significant performance degradation. Some visualized results shown in Fig. 9 demonstrate that rigid transformation is only effective in addressing perspective deformation while failing in other cases. In contrast, our non-rigid pre-alignment method proves to be more ro-

bust and effective in dealing with various situations.

Details recurrent refinement. To validate the effectiveness of this module, we replace it with another *Local correlation C_L & flow prediction* block as in Fig. 2 and retrain the model with the same settings. The output size for this variant is one-quarter of the input size, which will be bilinearly up-sampled. Results on the DocUNet dataset in Table 7 demonstrate the improvement introduced by details recurrent refinement.

5. Conclusions

We present DocAligner for automating the annotation of photographic document images by establishing dense correspondence between these images and their clean counterparts. DocAligner incorporates several sophisticated techniques, such as non-rigid pre-alignment, hierarchical alignment, details recurrent refinement, a synthesis pipeline, and a self-supervision training approach. Through extensive experiments, we demonstrate the excellent performance of DocAligner in the task of document dense correspondence, generating high-quality annotations. Moreover, DocAligner simplifies the manual annotation process to merely taking pictures, resulting in significant cost reduction associated with manual labeling. When annotating layout analysis data, DocAligner reduces the manual labeling time by a minimum of 30-fold. Experimental results in layout analysis, illumination correction, and geometric rectification emphasize the substantial potential of DocAligner in facilitating various document artificial intelligence tasks in photographic contexts. In our future work, we plan to create additional large-scale real-world datasets for the research community.

References

- [1] Hengyue Bi, Canhui Xu, Cao Shi, Guozhu Liu, Yuteng Li, Honghong Zhang, and Jing Qu. SRRV: A novel document object detector based on spatial-related relation and vision. *IEEE TMM*, 2022. 3
- [2] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE TPAMI*, 11(6):567–585, 1989. 2
- [3] Zewen Chi, Heyan Huang, Heng-Da Xu, Houjin Yu, Wanxuan Yin, and Xian-Ling Mao. Complicated table structure recognition. *arXiv preprint arXiv:1908.04729*, 2019. 1
- [4] Seokju Cho, Sunghwan Hong, and Seungryong Kim. Cats++: Boosting cost aggregation with convolutions and transformers. *IEEE TPAMI*, 2022. 4
- [5] Sagnik Das, Ke Ma, Zhixin Shu, Dimitris Samaras, and Roy Shilkrot. DewarpNet: Single-image document unwarping with stacked 3D and 2D regression networks. In *ICCV*, pages 131–140, 2019. 3, 8
- [6] S Das, H Ma Sial, R Baldrich, M Vanrell, and D Samaras. Intrinsic decomposition of document images in-the-wild. In *BMVC*, 2020. 3
- [7] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *ICCV*, pages 2758–2766, 2015. 4
- [8] Hao Feng, Yuechen Wang, Wengang Zhou, Jiajun Deng, and Houqiang Li. DocTr: Document image transformer for geometric unwarping and illumination correction. In *ACM MM*, pages 273–281, 2021. 3, 8
- [9] Hao Feng, Wengang Zhou, Jiajun Deng, Yuechen Wang, and Houqiang Li. Geometric representation learning for document image rectification. In *ECCV*, pages 475–492, 2022. 8
- [10] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2, 8
- [11] Kai Han, Rafael S Rezende, Bumsuk Ham, Kwan-Yee K Wong, Minsu Cho, Cordelia Schmid, and Jean Ponce. SC-Net: Learning semantic correspondence. In *ICCV*, pages 1831–1840, 2017. 2
- [12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *CVPR*, pages 2961–2969, 2017. 7
- [13] Sunghwan Hong and Seungryong Kim. Deep matching prior: Test-time optimization for dense correspondence. In *ICCV*, pages 9907–9917, 2021. 2
- [14] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, pages 2462–2470, 2017. 2
- [15] Sangryul Jeon, Seungryong Kim, Dongbo Min, and Kwanghoon Sohn. Parn: Pyramidal affine regression networks for dense semantic correspondence. In *ECCV*, pages 351–366, 2018. 3
- [16] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. COTR: Correspondence transformer for matching across images. In *ICCV*, pages 6207–6217, 2021. 4
- [17] Frédéric Kaplan, Sofia Ares Oliveira, Simon Clematide, Maud Ehrmann, and Raphaël Barman. Combining visual and textual features for semantic segmentation of historical newspapers. *Journal of Data Mining & Digital Humanities*, 2021. 3
- [18] Seungryong Kim, Stephen Lin, Sangryul Jeon, Dongbo Min, and Kwanghoon Sohn. Recurrent transformer networks for semantic correspondence. In *NeurIPS*, pages 6126–6136, 2018. 2
- [19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5
- [20] Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. TableBank: Table benchmark for image-based table detection and recognition. In *LREC*, pages 1918–1925, 2020. 1
- [21] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. DocBank: A benchmark dataset for document layout analysis. In *COLING*, pages 949–960, 2020. 1

- [22] Xiaoyu Li, Bo Zhang, Jing Liao, and Pedro V Sander. Document rectification and illumination correction using a patch-based CNN. *ACM TOG*, 38(6):1–11, 2019. 3, 8
- [23] Ruijiao Long, Wen Wang, Nan Xue, Feiyu Gao, Zhibo Yang, Yongpan Wang, and Gui-Song Xia. Parsing table structures in the wild. In *ICCV*, pages 944–952, 2021. 1
- [24] Ke Ma, Sagnik Das, Zhixin Shu, and Dimitris Samaras. Learning from documents in the wild to improve document unwarping. In *ACM SIGGRAPH*, pages 1–9, 2022. 2, 3, 6, 8
- [25] Ke Ma, Zhixin Shu, Xue Bai, Jue Wang, and Dimitris Samaras. DocUNet: Document image unwarping via a stacked U-Net. In *CVPR*, pages 4700–4709, 2018. 1, 6, 8
- [26] Iaroslav Melekhov, Aleksei Tiulpin, Torsten Sattler, Marc Pollefeys, Esa Rahtu, and Juho Kannala. DGC-Net: Dense geometric correspondence network. In *WACV*, 2019. 2, 4, 6, 7
- [27] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jae-heung Surh, Minjoon Seo, and Hwalsuk Lee. CORD: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS*, 2019. 1
- [28] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NeurIPS Autodiff Workshop*, 2017. 5
- [29] Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter WJ Staar. DocLayout: A large human-annotated dataset for document-layout analysis. In *ACM SIGKDD*, page 3743–3751, 2022. 1
- [30] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *CVPR*, pages 6148–6157, 2017. 2, 3
- [31] Xi Shen, François Fleuret, Alexei A Efros, and Mathieu Aubry. RANSAC-Flow: generic two-stage image alignment. In *ECCV*, pages 618–637, 2020. 2, 3, 4, 5, 6
- [32] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *NeurIPS*, volume 28, 2015. 2, 4
- [33] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *CVPR*, pages 8934–8943, 2018. 2, 4
- [34] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419, 2020. 2
- [35] Prune Truong, Martin Danelljan, Luc V Gool, and Radu Timofte. GOCor: Bringing globally optimized correspondence volumes into your neural network. In *NeurIPS*, pages 14278–14290, 2020. 2, 6, 7
- [36] Prune Truong, Martin Danelljan, and Radu Timofte. GLU-Net: Global-local universal network for dense flow and correspondences. In *CVPR*, pages 6258–6268, 2020. 2, 4, 6, 7
- [37] Xuan-Son Vu, Quang-Anh Bui, Nhu-Van Nguyen, Thi Tuyet Hai Nguyen, and Thanh Vu. MC-OCR challenge: Mobile-captured image document recognition for vietnamese receipts. In *RIVF*, pages 1–6, 2021. 1
- [38] Jiapeng Wang, Chongyu Liu, Lianwen Jin, Guozhi Tang, Jiaxin Zhang, Shuaítao Zhang, Qianying Wang, Yaqiang Wu, and Mingxiang Cai. Towards robust visual information extraction in real world: new dataset and novel solution. In *AAAI*, volume 35, pages 2738–2745, 2021. 1
- [39] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multi-scale structural similarity for image quality assessment. In *ACSSC*, volume 2, pages 1398–1402, 2003. 6
- [40] Guo-Wang Xie, Fei Yin, Xu-Yao Zhang, and Cheng-Lin Liu. Document dewarping with control points. In *ICDAR*, pages 466–480, 2021. 3, 8
- [41] Chuhui Xue, Zichen Tian, Fangneng Zhan, Shijian Lu, and Song Bai. Fourier document restoration for robust document dewarping and recognition. In *CVPR*, pages 4573–4582, 2022. 3, 6, 8
- [42] Xiao Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, and C Lee Giles. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. In *CVPR*, pages 5315–5324, 2017. 3
- [43] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE TPAMI*, 35(12):2878–2890, 2012. 6
- [44] Shaodi You, Yasuyuki Matsushita, Sudipta Sinha, Yusuke Bou, and Katsushi Ikeuchi. Multiview rectification of folded documents. *IEEE TPAMI*, 40(2):505–511, 2017. 6
- [45] Jiaxin Zhang, Canjie Luo, Lianwen Jin, Fengjun Guo, and Kai Ding. Marior: Margin removal and iterative content rectification for document dewarping in the wild. In *ACM MM*, pages 2805–2815, 2022. 2, 3
- [46] Peng Zhang, Can Li, Liang Qiao, Zhanzhan Cheng, Shiliang Pu, Yi Niu, and Fei Wu. VSR: a unified framework for document layout analysis combining vision, semantics and relations. In *ICDAR*, pages 115–130, 2021. 3
- [47] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. PubLayoutNet: largest dataset ever for document layout analysis. In *ICDAR*, pages 1015–1022, 2019. 1, 6

Appendix

A. Detailed network architecture

The architecture of our flow field decoder $\text{deocder}_l()$ described in Equation 3 is shown in Fig. 10, where the output channels for each convolution layer are 128, 128, 96, 64, 32, and 2, respectively. The skip connection will be removed for the lowest resolution level (i.e., $l = 1$).

The structure of our ConvGRU unit is depicted in Fig. 11. This unit takes an input x^n and a hidden state h^{n-1} , and outputs an updated hidden state h^n along with a residual flow Δf^n . The f -decoder and w -decoder in ConvGRU unit are 2-layer convolutional neural networks with a ReLU activation function. The outputs of f -decoder and w -decoder are respectively $\Delta f^n \downarrow \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 2}$ and a weight map $\in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 144}$. To obtain the up-sampled residual flow Δf^n , we first reshape the weight map to a size of $\frac{H}{4} \times \frac{W}{4} \times 16 \times 3 \times 3$, corresponding to 16 weight matrices of size 3×3 for each pixel in $\Delta f^n \downarrow$. Then, we calculate each up-sampled pixel value by performing a weighted sum over the 3×3 neighborhood in $\Delta f^n \downarrow$ using the 3×3 weight matrices from the weight map. Finally, we obtain the desired up-sampled residual flow Δf^n .

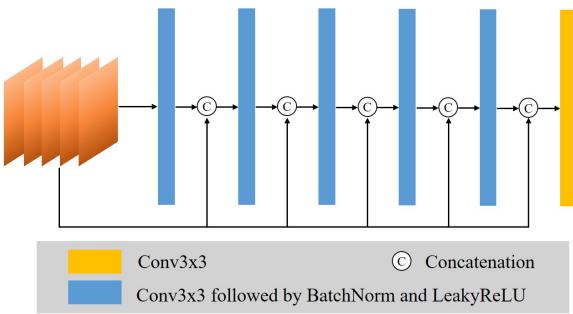


Figure 10. Architecture of our flow field decoders.

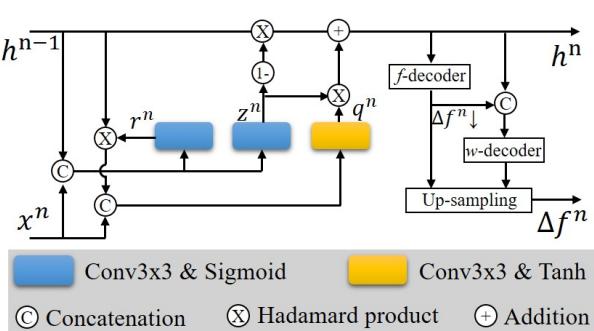


Figure 11. Architecture of our ConvGRU unit.

B. Optimization during self-supervision learning

As described in the main text, we estimate the solution of the energy function (Equation 10) using a gradient descent algorithm, rather than solving it directly. Specifically, we utilize Adam as our gradient descent optimizer with a fixed learning rate of 1×10^{-4} . For a more stable optimization, we perform augmentations by warping I'_s in each pre-aligned pair (I'_s, I_t) three times using different randomly-selected flow fields from DocAlign12K. The resulting augmented pairs, along with the original (I'_s, I_t) , form a mini-batch of size 4 for optimization during each iteration.

C. Details on synthesising DocAlign12K

Our PDF files sourced from magazines, textbooks, scientific articles, and handwritten notes, among others, are predominantly in Chinese and English. We convert these PDF files into images and treat them as target clean document images.

We initialize a flow field with a size of $1024 \times 1024 \times 2$, and then assign a 2-dimensional vector (dx_i, dy_i) for each pixel i . Values of dx_i and dy_i are randomly selected from $[-4096, 4096]$ and indicate the direction and distance the pixel is to be shifted. We ensure the flow field is continuous and smooth by applying double mean filtering with a kernel size of 91 each. Although the flow field now contains local perturbations, it still lacks global deformations. Therefore, we add global distortion by considering translation and scaling. Random translation is achieved by respectively adding a random value from $[-50, 50]$ to x and y channel in the flow field. For random scaling, we create a base coordinate map $base_coordinate_{ij} = (i, j) \in \mathbb{R}^{1024 \times 1024 \times 2}$ and a scaling map $scaling_{ij} = (scaling_x, scaling_y) \in \mathbb{R}^{1024 \times 1024 \times 2}$. The values for $scaling_x$ and $scaling_y$ are chosen randomly from $[-0.05, 0.2]$. We introduce the scaling deformation to the flow field with the following equation:

$$f = f + (base_coordinate - 512) \otimes scaling, \quad (13)$$

where \otimes denotes the Hadamard product and f is our generated flow field. Finally, we use the resulting 2-dimensional flow field to sample geometrically-distorted document images from clean targets. The visualization of our sampled results are shown in Fig. 12.

D. Additional results for DocAligner's application on layout analysis

We present additional visualization results for the 2200 samples we acquired for document layout analysis (DLA) in Fig. 13. These results demonstrate that our DocAligner can produce photographic DLA data that includes various



Figure 12. Geometrically-distorted document images resulting from our randomly-generated flow fields.

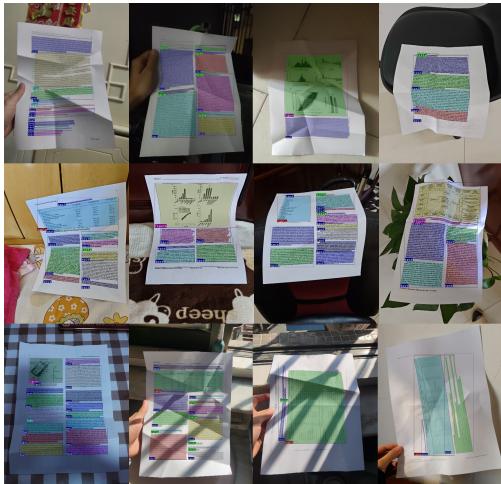


Figure 13. More visualized results for our acquired DLA dataset.

illuminations, geometric deformation, environmental margin, and document content. With high-precision pixel-level alignment, DocAligner can generate accurate and tight-fitting masks and bounding boxes even for geometrically-distorted documents. This is an enormous advantage over manual labeling, which requires a significant amount of time and effort.

As described in our main text, we train a Mask R-CNN with ResNet-50 backbone using our collected 2000 training data. Our anchor aspect ratios are set to 0.5, 1, 2, 4, 8, 16, and 24. We train the model on two NVIDIA 2080Ti GPUs with a batch size of 4. We then use the resulting model to inference on 200 test images. Visualized results in Fig. 14 demonstrate that Mask-RCNN trained with DocAligner-acquired data produces suitable performance with such a small scale. This indicates that the acquired data is effective and valuable for further research and development in DLA.



Figure 14. More visualized results for our trained Mask R-CNN.

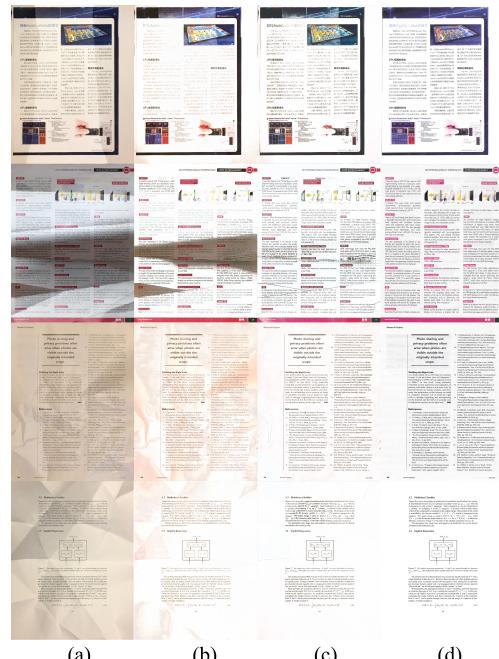


Figure 15. From left to right, figures illustrate (a) input images, (b) results from the model trained with the DocProj synthetic dataset, (c) results from the model trained with real-world dataset acquired by DocAligner, and (d) scanned ground-truth.

E. Additional results for DocAligner’s application on illumination correction

As mentioned in Section 4.4, we have trained illNet using the synthetic dataset, DocProj, and the real-world dataset, which were acquired by DocAligner. Visualized results are shown in Fig. 15. We observe that the model trained using our acquired real-world dataset is more effec-

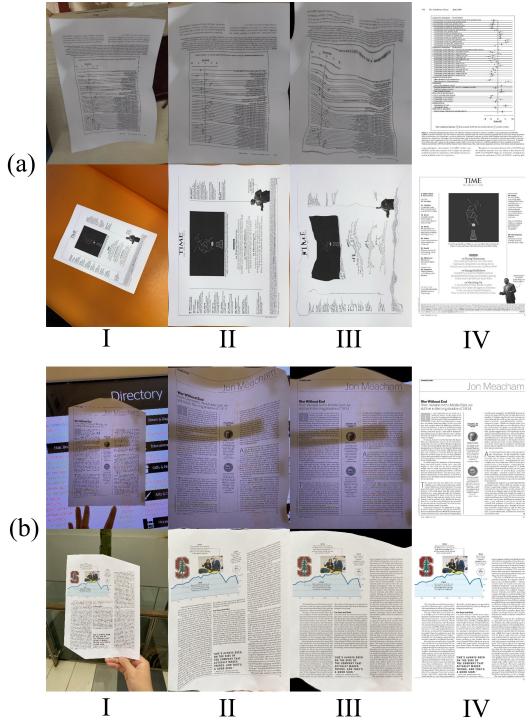


Figure 16. Several failure cases encountered during the application of DocAligner. Specifically, (a) illustrates the issue of orientation misalignment, while (b) depicts incomplete documents. From left to right, the figures show I the photographics, II pre-aligned results , III final aligned results, and IV target images.

tive in preserving the textual details and reducing the impact of shadows and creases, even though the size of the real-world dataset is relatively small.

F. Limitations

The current implementation of DocAligner has some limitations, which need to be considered. Firstly, the orientation of photographic images needs to be correct, as our non-rigid pre-alignment module is not orientation aware. Fig. 16 (a) illustrates the problematic results produced by DocAligner in pairs containing orientation misalignment. Hence, controlling the camera’s attitude to determine the correct orientation is essential. Alternatively, manual adjustment of orientation after capturing is needed.

Secondly, when the source document is incomplete, there may be invalid areas in the aligned results which we fill with zeros. As depicted in Fig. 16 (b), such areas make the aligned results unsuitable for document illumination correction tasks.