

Supplementary materials

The supplementary materials contain five sections:

- A: Details of the CRN Architecture
- B: Details of Experimental Setting
- C: More Visual Comparisons on Document Registration
- D: More Experimental Results of Document Dewarping

A DETAILS OF THE CRN ARCHITECTURE

Figure 9 shows our CRN (coarse registration network) described in Sec. 3.1. In general, this architecture is an extension of the dewarping network, where we add a flat target image I_t as prior information to form a pair of inputs. We first use two identical geometric heads to extract shallow features from the paired warped-flat document images. In detail, we adopt an EfficientNet B7 noisy student model [41] as the head to perform shallow feature extraction on the given input $I_{so} \in \mathbb{R}^{H \times W \times 3}$, obtaining $f_{so} \in \mathbb{R}^{35 \times 35 \times 128}$ and $f_t \in \mathbb{R}^{35 \times 35 \times 128}$ where $H = W = 280$. Then, the two sets of features are concatenated and fed into the visual transformer (ViT) to predict mapping flow $f_c \in \mathbb{R}^{H \times W \times 2}$, where we leverage a learnable upsampling module proposed in GeoTr [12]. Finally, we can apply f_c to sample I_{so} to obtain a coarse-dewarped image I_{sw} by the `grid_sample` function in PyTorch. We can see the warp of I_{sw} is mild compared to the original input document I_{so} , demonstrating the effectiveness of the CRN.

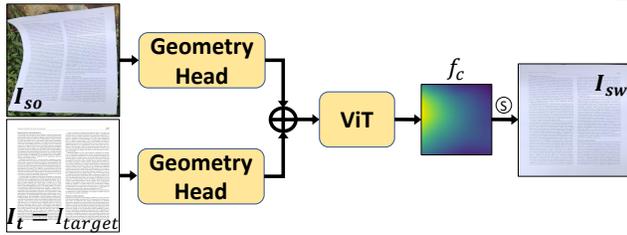


Figure 9: CRN (coarse registration network) Architecture

B DETAILS OF EXPERIMENTAL SETTING

To better understand our experiments, we summarize the details of our experimental setting in the training of CRN, Cross-reconstruction pre-training, and FRN in the following tables Tab.5, Tab.6, and Tab.7, respectively. For CRN, we basically refer to the optimal settings of Inv3D [17].

C MORE VISUAL COMPARISONS

To better demonstrate the effectiveness of our proposed document registration pipeline, we show more visual qualitative comparisons in Fig. 10, which is a supplement to Fig. 6. We can see that the existing methods either cannot align character-level texture features well, such as Fig. 10(b)(d), or they are prone to produce error matches in texture-less areas such as Fig. 10(c). Our registration method can effectively avoid these issues and is suitable for various warping patterns (such as curves and folds), thus obtaining better results, as shown in Fig. 10(e). Specifically, compared to Fig. 10(b)(d),

Table 5: Experimental Setting for CRN

Settings	Value
training data	Doc3D
data augmentation	Random crop
ViT layers	6
embedding dimension	256
ViT patch size	8×8
input resolution	280×280
training epochs	300
batch size	8
optimizer	AdamW
scheduler	OnceCycleLR
initial learning rate	4E-6
peak learning rate	1E-3

Table 6: Experimental Setting for Cross-Reconstruction Pre-training

Settings	Value
training data	coarse-dewarped image I_{sw} and flat counterpart I_t
data augmentation	ColorJitter, RandomGrayscale
ViT patch size	16×16
ViT layers	11
input resolution	384×512
training epochs	100
warm-up epochs	10
batch size	12
optimizer	AdamW
scheduler	Cosine
initial learning rate	4E-6
peak learning rate	1E-3
EMA coeff	0.995
weight decay	5E-2

Table 7: Experimental Setting for FRN

Settings	Value
training data	coarse-dewarped image I_{sw} and flat counterpart I_t
data augmentation	Shadow Replacement
ViT patch size	16×16
ViT decoder layers	5
embedding dimension	64,128,256,512,1024
input resolution	384×512
training epochs	80
batch size	12
optimizer	AdamW
scheduler	MultiStepLR
learning rate milestones	30,60 by factor 0.5
initial learning rate	2E-5 and 4E-4
	for encoder,decoder respectively

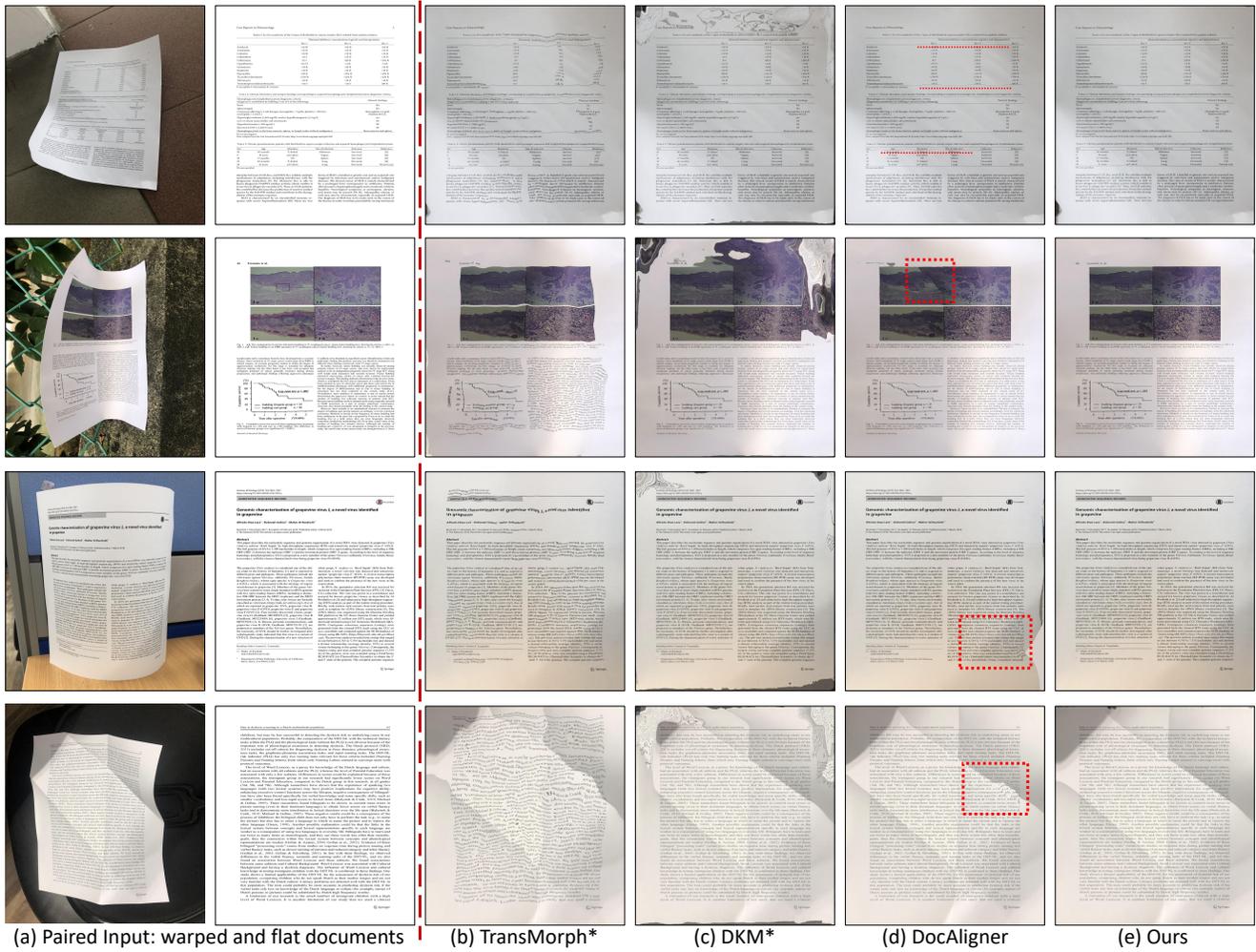


Figure 10: More Qualitative comparison on the proposed WarpDoc-R benchmark. Our method is superior to existing registration methods(b~d) in both character-level alignment and texture-less areas. Zoom-in is recommended for better visualization. Since TransMorph and DKM are originally proposed for general images, we re-trained them on document datasets and specified by the symbol "*".

our method can obtain better registration results at different scales. For large-scale objects, the horizontal table lines (red dotted dashes in the first row) registered by ours are more horizontal, and the 2×2 figure (in the second row) maintains a regular rectangular structure. In the case of the third and fourth rows, we exhibit some small-scale objects; we can register fine-grained characters without distortion, while Fig. 10(d) is prone to local distortion (red dotted box part). Compared with Fig. 10(c), our method greatly suppresses error matching in texture-less areas, which we believe is mainly attributed to the design of the location classifier mentioned in Sec. 3.3.

D MORE EXPERIMENTAL RESULTS OF DOCUMENT DEWARPING

To better verify the effectiveness and high quality of the registered real document data, we evaluate the dewarping performance by

Table 8: Dewarping quantitative comparisons on the dewarping model DewarpNet [8] trained by different dataset scales.

Dataset	Data type	Dataset scale	MS-SSIM \uparrow	LD \downarrow	AD \downarrow	ED \downarrow	CER \downarrow
Doc3D [8]	Synthetic	0.8k	0.441	16.7	0.375	992	0.263
	Synthetic	20k	0.443	14.8	0.273	628	0.175
Ours	Synthetic	80k	0.464	12.3	0.236	516	0.142
	Real	0.8k	0.520	10.8	0.161	441	0.131
Ours+Doc3D	Real+Synthetic	0.8+20k	0.516	11.5	0.197	468	0.138

another representative dewarping model DewarpNet [8]. Similar to Tab. 2, we conduct experiments on synthetic and real documents with different dataset scales. As shown in Tab. 8, given only registered 800 real document training samples, we can exceed the dewarping performance of $100 \times$ dataset scale of synthetic training data. The similar phenomenon in both models (Tab. 2 and Tab. 8)



Figure 11: Qualitative visual comparison of the dewarping model DewarpNet [8] on WarpDoc-R benchmark and the synthetic data Doc3D with different number of dataset scales.

illustrates the effectiveness of our registration pipeline. Besides, the results in the last row of Tab. 8 also confirm that the mixture of a small number of real and a large number of synthetic documents benefits the model to obtain better performance. This again confirms the insight that *fewer high-quality real samples are more valuable than a large amount of low-quality synthetic data*.

Furthermore, we add more visual dewarping comparison results as a supplement to Fig. 7, as shown in Fig. 11. We can see that our model trained on 800 registered real documents (Fig. 11(c)) surpasses all models trained on purely synthetic data with different dataset scales including $1\times$ scale (Fig. 11(b)), $25\times$ scale (Fig. 11(d)) or $100\times$ scale (Fig. 11(e)). It can be seen from Fig. 11(b)(d)(e) that with the amount of synthetic data gradually increasing, the difference in visual results of dewarping has only a small improvement, such as background removal. Even some results may introduce additional distortion. For instance, 80k in the third row introduces greater paragraph distortion than 20k. We argue this may be due to the model’s tendency to overfit on limited synthetic data. On the contrary, we can achieve more satisfactory results with only 0.8k samples on real

document images, which confirms the effectiveness of the proposed registration pipeline and the significance of using registered real documents for dewarping training. Moreover, extra performance improvements can be achieved by mixing real(0.8k) and synthetic (20k) data, as shown in Fig. 11(f). We hope that these phenomena will inspire more research to explore how to mix synthetic and real data better to aid training.