Marior: Margin Removal and Iterative Content Rectification for Document Dewarping in the Wild

Jiaxin Zhang  
Canjie Luo  
South China University of Technology  
msjxzh@mail.scut.edu.cn  
canjie.luo@gmail.com

Lianwen Jin  
South China University of Technology  
Peng Cheng Laboratory  
Pazhou Laboratory (Huangpu)  
eelwjin@mail.scut.edu.cn

Fengjun Guo  
Kai Ding  
IntSig Information Co. Ltd  
fengjun_guo@intsig.net  
danny_ding@intsig.net

Figure 1: Cases that have not been sufficiently studied by existing deep learning-based methods. (a) Document image with large marginal regions. (b) Document image without marginal regions. Marginal region refers to the area composed of pixels that do not belong to the document of interest. We highlight the deformation using red dotted lines.

ABSTRACT

Camera-captured document images usually suffer from perspective and geometric deformations. It is of great value to rectify them when considering poor visual aesthetics and the deteriorated performance of OCR systems. Recent learning-based methods intensively focus on the accurately cropped document image. However, this might not be sufficient for overcoming practical challenges, including document images either with large marginal regions or without margins. Due to this impracticality, users struggle to crop documents precisely when they encounter large marginal regions. Simultaneously, dewarping images without margins is still an insurmountable problem. To the best of our knowledge, there is still no complete and effective pipeline for rectifying document images in the wild. To address this issue, we propose a novel approach called Marior (Margin Removal and Iterative Content Rectification). Marior follows a progressive strategy to iteratively improve the dewarping quality and readability in a coarse-to-fine manner. Specifically, we divide the pipeline into two modules: margin removal module (MRM) and iterative content rectification module (ICRM). First, we predict the segmentation mask of the input image to remove the margin, thereby obtaining a preliminary result. Then we refine the image further by producing dense displacement flows to achieve content-aware rectification. We determine the number of refinement iterations adaptively. Experiments demonstrate the state-of-the-art performance of our method on public benchmarks. The resources are available at https://github.com/ZZZHANG-jx/Marior for further comparison.

CCS CONCEPTS

• Computing methodologies → Computer vision;  
• Applied computing → Document scanning; Document management and text processing.

KEYWORDS

Camera-captured document images, Dewarping, Geometric rectification, Convolutional neural network

ACM Reference Format:

Jiaxin Zhang, Canjie Luo, Lianwen Jin, Fengjun Guo, Kai Ding. 2022. Marior: Margin Removal and Iterative Content Rectification for Document Dewarping in the Wild. In Proceedings of the 30th ACM International Conference on Multimedia (MM ’22), October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3503161.3548214

1 INTRODUCTION

Powered by advanced built-in cameras in mobile devices, digitizing ubiquitous documents in daily life has become convenient for people. However, because of the inappropriate angle and position of the camera, the captured document images usually contain perspective...
deformation. Besides, the document itself may also geometrically deform because of curving, folding, or creasing. These types of deformation lead to the deteriorated performance of optical character recognition (OCR) systems and poor readability for readers.

Recent deep learning-based dewarping methods [7, 9–13, 23, 25, 29, 40, 41] have made great progress regarding robustness to a variety of document layouts. However, almost all of them only focus on accurately cropped document images, and ignore cases with large marginal regions or without marginal regions, which are shown in Fig. 1 (a) and (b), respectively. In this study, marginal region refers to the area composed of pixels that do not belong to the document of interest. To address this issue, one can take all these situations into consideration during training, but we found the results unsatisfactory (refer to supplementary material). We argue that this is attributed to the extra implicit learning to identify the foreground document and remove marginal region, which is also observed in [11]. Another way is to implement an existing object detection algorithm ahead of dewarping to avoid the need for manually cropping. However, dewarping document images without margins is still an unresolved problem. Accordingly, there is still no complete and effective pipeline for handling all cases in the wild.

Therefore, we propose Marior (Margin Removal and Iterative Content Rectification) to tackle this problem, which consists of two cascaded modules: margin removal module (MRM) and iterative content rectification module (ICRM). Marior decouples the margin removal and document rectification process. Specifically, in the MRM, we first feed the source distorted image into our mask prediction network, which predicts the corresponding document segmentation mask. Then we propose a novel mask-based dewarper (MBD) to remove the margin based on this mask and obtain a preliminary dewarped result. For images without marginal regions and no complete document edges, as shown in Fig. 1 (b), we propose filtering them out and skipping the margin removal process by using an intersection-over-union (IoU)-based method, which is inspired by the observation that these images usually result in noisy masks.

Thereafter, we feed the margin-removed output from the MRM into the ICRM for further refinement. It predicts a dense displacement flow that assigns a two-dimensional (2D) offset vector to each pixel in the input image. After rectification according to this flow, we obtain a dewarped output image. Because the margin-removed image focuses more on content (e.g., text lines and figures), the ICRM should be content-aware. Therefore, we further design a new content-aware loss to implicitly guide the ICRM to focus more on informative regions, such as text lines and figures, than the uniform document background. This design is based on an intuition that the latter contains fewer deformation clues and slight deviation for it on dewarped result is visually negligible. In addition, we have found that iterative implementation of ICRM can improve rectification performance. To this end, we propose an adaptive approach to determine the number of iterations to make the proposed iterative ICRM process more intelligent and efficient.

To summarize, our contributions are as follows:

- We propose a new method called Marior to handle document images that have various margin situations, which are ignored by existing learning-based methods.
- We propose a new mask-based dewarper in our margin removal module (MRM) that coarsely dewarps the document image based on the predicted segmentation mask. Then an iterative content rectification module (ICRM) is proposed to further refine the image by predicting a dense displacement flow.
- We design a new content-aware loss to implicitly guide the flow prediction network to focus more on informative regions. We also propose an adaptive iteration strategy to improve performance.
- Extensive experiments demonstrate that the proposed Marior achieves state-of-the-art performance on two widely used public benchmarks. Moreover, this approach achieves significant success for tackling cases that have diverse margins.

2 RELATED WORK

2.1 Traditional hand-crafted methods

Document image dewarping, or geometric rectification, has been widely studied in the literature. Most traditional hand-crafted methods attempted to reconstruct the three-dimensional (3D) shape of document images. The most direct approach is to use additional equipment, such as a depth sensor [45], visible light projector-camera system [2], and structured laser light sources [27]. This type of method can achieve significant performance in specific scenarios, but its application is very limited. Methods in [18, 35, 42] used multiple images from different views to reconstruct the 3D shape rather than relying on additional equipment. However, they are not sufficiently practical because multiple images are not always available in many situations. To reconstruct the 3D shape from a single image without additional equipment, researchers used low-level visual cues in document images, such as illumination effects [33, 44], text lines [22, 34], and boundaries [3, 4]. Such low-level visual cues were also widely used in some 2D image processing-based methods [14, 15, 32, 46] that did not need to reconstruct the 3D shape. Methods that relied on these low-level visual cues made strong assumptions about the document layout or deformation type, which were not sufficiently robust for real-world scenes.

2.2 Learning-based methods

Learning-based methods are becoming increasingly popular because of the availability of large datasets [1, 7, 25]. Das et al. [8] used convolutional neural networks (CNNs) in this task for the first time, where they only adopted CNNs to detect creases to help subsequent 2D image processing-based methods. More studies have emerged since the proposal of two synthetic datasets [7, 25]. These two datasets contain pixel-wise unwarping map annotation that can be directly used to sample the dewarped image from the source distorted image. Ma et al. [25] used a stacked U-Net to obtain a refined forward map. Das et al. [7] used several pixel-wise annotations from their Doc3D dataset to train CNNs to predict the 3D coordinates, backward map, and shading map, which were used to dewarp the image and adjust the illumination. Liu [23] proposed a multi-resolution method and a generative adversarial network (GAN) framework to make the output more visually pleasing. Li et al. [20] proposed a patch-based method to split the input image into small patches both during training and inference, thus reducing
As shown in Fig. 2, Marior contains two cascaded MBD and ICRM modules to alleviate that problem and improve performance. Amir et al. [26] explored the document content to guide the optimization of the neural network. Feng et al. [10] used the Transformer [36] as network architecture and predicted a dense displacement flow that has the same resolution as \( I_{pd} \). This mask-based dewarping process was very time-consuming. Das et al. [9] proposed an end-to-end trainable patch-based method to alleviate that problem and improve performance. Amir et al. [26] explored the content and achieved superior performance. However, these methods can only handle accurately cropped document images. There is still no complete and effective pipeline that is sufficiently robust to dewarp document images in the wild. Even for accurately cropped cases, the dewarping performance of these methods are still subpar for practical applications.

### 3 METHODOLOGY

As shown in Fig. 2, Marior contains two cascaded MBD and ICRM modules which progressively rectify the distorted source image \( I_s \) and output the final dewarped image \( I_{fd} \). In the MRM, we first remove the margin based on the predicted mask and obtain a preliminary dewarped result \( I_{pd} \). This mask-based dewarping process is achieved using a novel MBD. Then the ICRM takes as input \( I_{pd} \) and predicts a dense displacement flow that has the same resolution as \( I_{pd} \). This 2D flow assigns the distance that each pixel in \( I_{pd} \) should be shifted to obtain \( I_{fd} \). We sample \( I_{fd} \) from \( I_{pd} \) based on this flow. To gain better rectification performance, we implement the ICRM iteratively. We propose an adaptive method to determine the number of iterations.

#### 3.1 Margin removal module (MRM)

**Mask prediction.** To remove the margin from a given image, we first localize document regions. We consider the localization as a semantic segmentation task, which aims to produce a mask that precisely represents document regions. The architecture of our mask prediction network is shown in Fig. 3 (a). We directly adopt the encoder and decoder from DeepLabv3+ [5]. In addition to the document mask, we design a head to produce an edge mask for auxiliary training. Furthermore, we observe that the document mask has a unique and relatively fixed pattern, such as relatively straight edges, one large connected area, and a shape close to a quadrilateral. We impose this prior knowledge in the MRM using a GAN framework as shown in Fig. 3 (a). We find this can effectively reduce noise on the produced mask, as shown in Fig. 3 (b).

The objective is defined as

\[
\min_{\text{MRM}} L_{\text{MRM}} = \min (\lambda L_{\text{prior}} + L_{\text{mask}} + L_{\text{edge}}),
\]

\[
\max_{\text{Dnet}} L_{\text{Dnet}} = \max L_{\text{Dnet}} - L_{\text{prior}}. 
\]

\[ L_{\text{mask}} = -\frac{1}{N} \sum_{i} [m_{di} \cdot \log (\hat{m}_{di})] + (1 - m_{di}) \cdot \log (1 - \hat{m}_{di})], \]

\[ L_{\text{edge}} = -\frac{1}{N} \sum_{i} [m_{ei} \cdot \log (\hat{m}_{ei})] + (1 - m_{ei}) \cdot \log (1 - \hat{m}_{ei})], \]

where \( \hat{m}_{di} \) and \( \hat{m}_{ei} \) denote the predicted classification for the \( i \)-th element in document mask \( \hat{M}_{d} \) and edge mask \( \hat{M}_{e} \), and \( m_{di} \) and \( m_{ei} \) are their corresponding ground truths, respectively. \( N \) is the number of elements in \( \hat{M}_{d} \). \( L_{\text{prior}} \) is a standard objective in GAN framework that guides the distribution of \( \hat{M}_{d} \) closer to that of ground-truth mask \( M_{d} \), and \( \lambda \) is the weight for this term:

\[
L_{\text{prior}} = E \left[ \log D_{\text{net}} (M_{d}') \right] + E \left[ \log (1 - D_{\text{net}} (\hat{M}_{d})) \right].
\]

Inspired by [24], we replace the one-hot coding of \( M_{d} \) with \( M_{d}' \). We denote the \( i \)-th element in \( M_{d}' \) by

\[
m_{d,i}' = \begin{cases} 
0.9, & \text{if } i = 1, 2, \ldots, N; \\
0.1, & \text{otherwise}.
\end{cases}
\]
coding positive sample and the generated negative sample under when optimizing the discriminator. It is worth noting that this mask prediction model can also be other alternative segmentation models, which just need to provide the segmentation mask of document regions.

Mask-based dewarper (MBD). After obtaining the document mask, we propose a new MBD to remove the margin and perform preliminary dewarping, as shown in Fig. 4. Specifically, based on the predicted mask, we first detect the four corners using the Douglas-Peucker algorithm [31] and then determine the order (left top, right top, right bottom, and left bottom) based on their relative positions. Then we can determine equidistant points on each edge (in our experiment, in addition to the four corners, we use three equidistant points on each edge). We match these control points to the corresponding positions of a rectangle. Then we use these key point pairs to perform thin plate spline (TPS) [38] interpolation on $I_s$, thereby removing the margin and obtaining $I_{pd}$. Notably, for document images without marginal regions, which do not have complete edges, as shown in Fig. 1 (b), we skip TPS interpolation and consider the original $I_s$ as the output of the MBD. We filter these images out by calculating the IoU between $\hat{M}_d$ and the mask derived from all detected control points, and setting a threshold for this IoU. This is inspired by the observation that document images without complete edges usually lead to noisy $\hat{M}_d$, thus resulting in relatively low IoU.

3.2 Iterative content rectification module (ICRM)

The results from preliminary dewarping using the MRM are not perfect. The reasons are twofold. The first reason is that the selection of equidistant points on each edge does not consider depth information; hence, this equidistant division is inconsistent with that performed on physical paper. The second reason is that, sometimes, the predicted mask is not sufficiently accurate when it encounters unclear edges or very complex margins. Additionally, document images without marginal regions skip preliminary dewarping, thereby still being untouched.

To further rectify $I_{pd}$, we propose the ICRM, which takes $I_{pd}$ as input and produces a dense displacement flow $\hat{D}$. We adopt the commonly used encoder-decoder with skip connections as our displacement flow prediction network. We adopt the attention strategy [39] in bottleneck and dilated convolution [43] to enlarge receptive filed to capture global information. As mentioned previously, the rectification of informative regions, such as text lines and figures, is intuitively more crucial than that of the uniform document background. We use the document content mask $M_c$ to design our content-aware loss $L_c$, which implicitly guides the network to focus more on informative regions. We also adopt shift invariant loss $L_s$ as in [25]. The final training loss of ICRM is expressed as

$$L_{ICRM} = L_c + \alpha L_s,$$

(7)

$$L_c = \frac{1}{N} \sum_{i} \left( \left\Arrowvert \hat{d}_i - d_i \right\Arrowvert + \beta \cdot m_{ci} \cdot \left\Arrowvert \hat{d}_i - \hat{d}_j \right\Arrowvert \right),$$

(8)

$$L_s = \frac{1}{2N^2} \sum_{i,j} \left( \left( \hat{d}_i - d_i \right) - \left( \hat{d}_j - d_j \right) \right)^2,$$

(9)

where $\hat{d}_i$, $d_i$, and $m_{ci}$ denote the $i$-th element in predicted displacement flow $\hat{D}$, ground-truth displacement flow $D$, and document content mask $M_c$, respectively. $\alpha$ and $\beta$ are constant weights.

Because we complete margin removal in the MRM, the ICRM is supposed to focus on content rectification without extra implicit learning to identity the foreground document and remove marginal region. The separation of margin removal also makes the ICRM capable of adopting an iterative scheme to rectify the document step by step, which we find can improve rectification performance. If margin removal is not decoupled, the network might learn to rectify the document based on document edges and tend to find them in every iteration, even if they do not exist, which will result in problematic output. Our iterative scheme is shown in Fig. 5 (a). In the beginning, we feed $I_{pd}$ into the displacement flow prediction network and obtain the first displacement flow $\hat{D}^1$, which we can then use to sample $I_{pd}^1$ from $I_{pd}$.
We implement our model in the PyTorch framework [28] and train it on a single NVIDIA 2080Ti GPU with a batch size of 8. We adopt Adam [17] as our optimizer, with a weight decay of $5 \times 10^{-4}$, the initial learning rate is set to $1 \times 10^{-4}$, which is reduced by a factor of 0.5 after every 5 epochs. We train both networks for 50 epochs and select the model with the best performance on the validation set as the final model. We empirically set $\alpha$, $\beta$, the IoU threshold for skipping the TPS interpolation, and $\tau$ to 0.005, 3.0, 96, and 60, respectively.

### 4.4 Comparison on public benchmarks

We consider vanilla DeepLabv3+ [5] without data augmentation as the baseline and present the improvement we obtained in Table 1. We validate models on the dataset proposed in [16], which consists of 120 real-world document images. This dataset is constructed for document localization and only annotated with four corners of the document, which we use to generate the quadrilateral ground-truth mask (these document images only contain perspective deformation). As shown in Table 1, data augmentation greatly improves the performance. The mask prediction network in our MRM also obtains improvement. The effectiveness of the introduction of the prior knowledge can be seen in Fig. 3 (b).

We further evaluate the effectiveness of our proposed content-aware loss on our Doc3D validation set. We use structural similarity index (SSIM) to evaluate the quality of the rectified image that resulted from $\bar{D}$. As shown in Table 2, we achieve the best image quality with the setting of $\beta = 3$, which indicates the contribution of our proposed content-aware loss.
Table 3: Quantitative comparisons on the DocUNet benchmark dataset [25]. “†” indicates results are from publicly available images or models, the sources of which are shown in the footnotes. “↑” and “↓” indicate that the higher the better and the lower the better, respectively. “¬” indicates that results or models are not publicly available. Bold and underline denote the best and the second-best results, respectively.

| Method              | Crop | Origin |
|---------------------|------|--------|
|                     | MS-SSIM ↑ | LD ↓ | CER (%) ↓ | MS-SSIM↑ | LD ↓ | CER (%) ↓ |
| DocUNet [25]        | 0.41  | 14.08 |   \   |   \   |   \   |   \   |
| AGUN [23]           | 0.4491 | 12.08 |   \   |   \   |   \   |   \   |
| RectiNet²† [1]      | 0.415 | 13.2 | 35.95±20.4 | 0.3847  | 14.34 | 46.39±24.3 |
| DewarpNet³† [7]     | 0.4693 | 8.98 | 21.68±20.0 | 0.4341  | 10.22 | 30.44±22.4 |
| Xie et al.⁴† [40]   | 0.4361 | 8.50 | 76.35±31.7 | 0.1868  | 20.75 | 81.78±29.4 |
| DocProj²† [20]      | 0.4071 | 11.46 | 35.95±27.0 | 0.2948  | 23.13 | 47.15±26.5 |
| PiecewiseUnwarp [9] | 0.4879 | 9.23 | 30.01±14  |   \     |   \   |   \   |
| DocTr⁶† [10]        | 0.5085 | 8.38 | 18.05±18.4 |   \     |   \   |   \   |
| Marior w/o ICRM     | 0.4380 | 9.56 | 25.88±19.9 | 0.4178  | 10.18 | 29.42±22.1 |
| Marior w/o iteration| 0.4659 | 8.15 | 20.86±19.4 | 0.4391  | 9.09  | 25.21±20.0 |
| Marior              | 0.4733 | 8.08 | 18.35±17.86 | 0.4458  | 9.00  | 22.34±19.7 |

(CER), which derives from the Levenshtein distance [19] between the recognized and reference text. The CER can be computed as \( CER = (s + i + d)/N \), where \( s, i, \) and \( d \) are the number of substitutions, insertions, and deletions from the Levenshtein distance, respectively. \( N \) is the number of characters in the reference text.

**DocUNet benchmark** [25]. The quantitative results on this dataset are given in Table 3, where “Crop” represents accurately cropped images that are usually used for comparison in previous studies. “Origin” represents originally captured images without cropping, thereby containing large marginal regions. For a more fair comparison, Faster R-CNN [30] is used as a document detector attached to other methods when conducting experiments on “Origin” subset. The details of this detector are included in supplementary material. Text recognition is performed on 50 text-rich images or models, the sources of which are shown in the footnotes. ↑ indicates that results or models are not publicly available. Bold and underline denote the best and the second-best results, respectively.

Table 4: OCR evaluation results after rectification using different methods. DL denotes the abbreviation of deep learning and Reg. denotes recognizer. Bold and underline denote the best and the second-best results, respectively.

| Method              | CER (%) ↓ | Time(s) |
|---------------------|-----------|---------|
|                     | Non-DL Reg. | DL Reg. |
| Source Image        | 16.12      | 15.29   | -       |
| RectiNet²† [1]      | 26.67      | 27.57   | 1.21    |
| DewarpNet³† [7]     | 7.01       | 5.46    | 0.85    |
| Xie et al.⁴† [40]   | 65.60      | 11.35   | 2.01    |
| DocProj²† [20]      | 6.13       | 3.12    | 0.08    |
| DocTr⁶† [10]        | 4.31       | 3.69    | 2.37    |
| Marior w/o ICRM     | 12.26      | 11.54   | 0.17    |
| Marior w/o iteration| 7.92       | 7.19    | 0.30    |
| Marior              | **4.18**   | **3.28** | **0.85** |

When compared to existing methods on the “Crop” subset, Marior achieves comparable performance. However, on the “Origin” subset, our method outperforms existing methods by a large margin even Marior is without the help of the detector. The qualitative comparisons are shown in Fig. 6 and 7. In Fig. 6, we compare our method with DocProj [20], DewarpNet [7], and the method of Xie et al. [40]. The input images in the first three columns are from the “Crop” subset. Although DocProj [20] rectifies the document content to some extent, the margin remains, which results in poor visual aesthetics. DewarpNet [7] and method of Xie et al. [40] rectify the document content well and simultaneously remove the margin. Our method also achieves good perceptual performance and performs better for details than the methods in [7] and [40]. The input images
Figure 6: Qualitative comparisons on the DocUNet benchmark dataset [25]. (a) Input, (b) DocProj [20], (c) DewarpNet [7], (d) method of Xie et al. [40], (e) Marior (ours), and (f) scanned ground truth. Input images in the first three columns are from the “Crop” subset and input images in columns 4 and 5 are from the “Origin” subset. Input images in columns 6 and 7 are obtained by further cropping the images in the “Crop” subset. We highlight the deformation using red dotted lines.

in the 4th and 5th columns are from the “Origin” subset, for which previous methods can achieve plausible results when with the help of a powerful document detector. By comparison, Marior can handle this subset without detector. As for input images without marginal regions in the 6th and 7th columns, Marior still achieves satisfactory performance, whereas existing methods do not. We make further comparisons with state-of-the-art methods PiecewiseUnwarp [9] and DocTr [10] in Fig. 7, which also demonstrates the superiority of our Marior.

OCR_REAL dataset [1]. This dataset contains text ground truth, which we consider as the reference text for CER metric. Besides, because of the lack of scanned ground-truth images, we do not evaluate the MS-SSIM and LD. Recognition performance is highly relative to the recognition engine. Therefore, to be more
We also evaluate the average running time of different methods on this dataset. For a fair comparison, we keep the resolution of the output image the same (1024 × 960) for each method when we evaluate the running time, which will differ when sampled images are with different resolution. The results are represented in Table 4, which shows that DocProj [20], DocTr [10] and Marior achieve stable and superior performance under both recognition engines when compared to the rest methods. However, DocProj [20] and DocTr [10] is more time consuming than Marior. Additionally, as analyzed previously and shown in Fig. 8, DocProj [20] fails to achieve visual aesthetics as Marior because of it’s disability to remove the margin.

5 CONCLUSIONS
We propose a simple yet effective method, Marior, to dewarp document images in a coarse-to-fine manner. We adopt two cascaded modules to first remove the margin of the document image and then further rectify the content. The proposed Marior adaptively determines the number of iterations, thus achieving a trade-off between efficiency and performance. Our proposed method not only achieves state-of-the-art performance on the DocUNet [25] and OCR_REAL [23] benchmark datasets but also successfully tackles cases with large marginal regions and cases without marginal regions, which are less investigated in previous studies. This is a significant success in terms of dewarping documents in the wild.

In future work, it is worth exploring the end-to-end optimization of the two proposed modules to achieve better performance.

ACKNOWLEDGMENTS
This research is supported in part by NSFC (Grant No.: 61936003), GD-NSF (no.2017A030312006, No.2021A1515011870), and the Science and Technology Foundation of Guangzhou Huangpu Development District (Grant 2020GH17)
Appendix

A TRAINING DEWARPNET BY USING DATA WITH DIVERSE MARGINAL SITUATION

Figure 9: Training data with different marginal situations. (I) Tightly cropped document images that are used by DewarpNet. (II) Document images with large marginal region and (III) without margins that are ignored in DewarpNet.

To enable models to handle images with diverse marginal situations, the most intuitive way is to take all these situations into consideration during training. As only considering dewarping tightly cropped document images, DewarpNet [7] only takes images like Fig. 9 (I) as training data while ignoring images with large marginal region or without margins like Fig. 9 (II) and (III). Here we include data in Fig. 9 (II) and (III) to retrain DewarpNet [7]. Note that all experiments are based on the public available training code of DewarpNet [7] and the only difference between these experiments is the type of training data. Results are given in Table 5, from which we can see that taking all marginal situations into consideration during training does not result in satisfactory performance. This is mainly due to the difficulty in predicting 3D coordinates and backward mapping of document when the depth and scale of document have large variations. The model implicitly learns to identify the foreground document and remove marginal region. In contrast, segmentation is a more easy and straightforward task when countered with these variations, which is adopted in Marior. Marior decouple the margin removal process and make the displacement flow prediction network focused on content rectification.

Table 5: Quantitative results when training DewarpNet with different types of data. (I), (II) and (III) represent tightly cropped images, images with large marginal regions and images without margins as in Fig. 9, respectively.

| Training data | Crop | Origin |
|---------------|------|--------|
| I  | II  | III  | MS-SSIM↑ | LD↓ | MS-SSIM↑ | LD↓ |
| ✓  | ✓   | ✓    | 0.4637   | 9.20 | 0.2055   | 56.87 |
| ✓  | ✓   | -    | 0.4666   | 9.01 | 0.3842   | 17.98 |
| ✓  | -   | ✓    | 0.4405   | 10.56| 0.1631   | 45.84 |
| ✓  | ✓   | ✓    | 0.4372   | 10.41| 0.3644   | 20.18 |

B ARCHITECTURE OF THE DISPLACEMENT FLOW PREDICTION NETWORK

We adopt the commonly used encoder-decoder with skip connections as our displacement flow prediction network. We implement the attention strategy [39] in bottleneck and dilated convolution [43] to enlarge receptive field and capture global information. The detail architecture is shown in Fig. 10.

Figure 10: The architecture of our displacement flow prediction network.

C THE EFFECT OF DIFFERENT ITERATIONS

We explore the effect of different iterations when we exclude our adaptive termination method. Results shown in Fig. 11 demonstrate that our iterative strategy can progressively improve metrics in the first few iterations. However, as the iteration continues, performance of these two metrics no longer improve and even deteriorate. So we terminate iteration by using our adaptive strategy. We also give some qualitative results in Fig. 12.

Figure 11: Dewarping performance under different iterations. We mark the best with red dot.

D THE DETAILS OF IMPLEMENTING DOCUMENT DETECTOR

We adopt Faster R-CNN [30] as our document detector, which is trained by using Doc3D [7] dataset. The ground-truth bounding box derives from the minimum bounding rectangle of the document mask. The resulting model achieves 98.9% average precision...
with IoU threshold of 75% (AP75) on dataset proposed in [16]. We apply this detector to crop the document out when we conduct experiments on the "Origin" subset. To be more consistent with the "Crop" subset, we enlarge the predicted bounding box by 30 pixels in all four directions. We consider the cropped image deriving from these bounding box as the input image for dewarping models. Results are given in Table 6. Marior without detector achieves better performance than other method with detector. When combined with the detector, Marior’s performance improve further.

Table 6: Results on the “Origin” subset. “***” denotes that the document detector is applied before dewarping. “†” indicates that results are from publicly available images or models.

| Method               | MS-SSIM↑ | LD ↓  | CER (%)↓ |
|----------------------|----------|-------|----------|
| RectiNet† [1]         | 0.2370   | 44.13 | 77.72±12.6 |
| RectiNet‡* [1]        | 0.3847   | 14.34 | 46.39±24.3 |
| DewarpNet† [7]        | 0.2022   | 53.16 | 70.98±26.0 |
| DewarpNet‡* [7]       | 0.4341   | 10.22 | 30.44±22.4 |
| Xie et al.† [40]      | 0.1001   | 42.16 | 99.28±1.9  |
| Xie et al.‡* [40]     | 0.1868   | 20.75 | 81.78±29.4 |
| DocProj† [20]         | 0.1799   | 52.97 | 87.54±22.2 |
| DocProj‡* [20]        | 0.2948   | 23.13 | 47.15±26.5 |
| Marior                | 0.4458   | 9.00  | 22.34±19.7 |
| Marior*               | 0.4666   | 8.38  | 20.22±19.6 |

E PERFORMANCE OF PREVIOUS METHODS ON IMAGES PRE-DEWARPED BY MRM

MRM can also be a pre-processing module for other methods. We give results of previous studies on images pre-dewarped by MRM as below, which demonstrate the superiority of ICRM over previous methods.

Table 7: Performance of previous methods on images pre-dewarped by MRM.

| Method                | MS-SSIM  | LD    | MS-SSIM  | LD    |
|-----------------------|----------|-------|----------|-------|
| MRM+RectiNet          | 0.4061   | 12.37 | 0.4024   | 12.46 |
| MRM+DewarpNet         | 0.4211   | 10.90 | 0.4186   | 10.95 |
| MRM+Xie et al.        | 0.1673   | 22.07 | 0.1824   | 21.71 |
| MRM+DocProj           | 0.2725   | 13.71 | 0.2785   | 13.66 |
| MRM+ICRM (Marior)     | 0.4733   | 8.08  | 0.4458   | 9.00  |