A Gated and Bifurcated Stacked U-Net for Document Image Dewarping (RectiNet)

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Hmrishav Bandyopadhyay (Presenter), Tanmoy Dasgupta, Dr Nibaran Das, Dr Mita Nasipuri

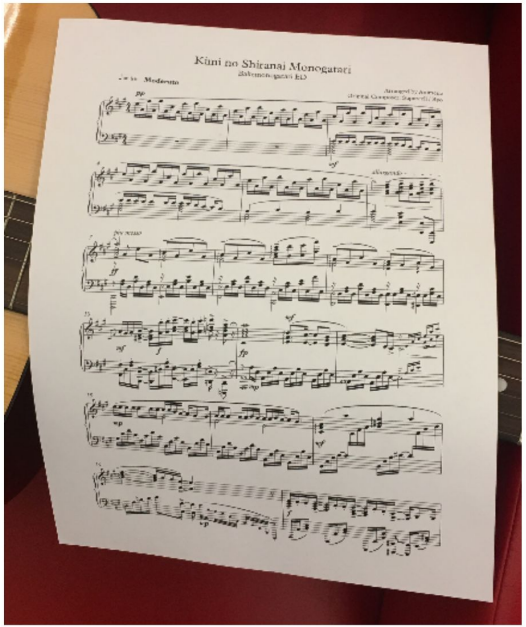
Jadavpur University

Domain of work

Document Dewarping Methods are necessary for

1. Unwarping digitized hard-copies of documents in an easy manner, and to help recognize what is written in those documents without human supervision.
2. OCR systems face difficulties when they are given texts which are at times not aligned or have some words which are out of alignment due to poor document structure in the image.

A pictorial representation



RectiNet
end-to-end network

Gated and
bi-furcated
Stacked U-Net

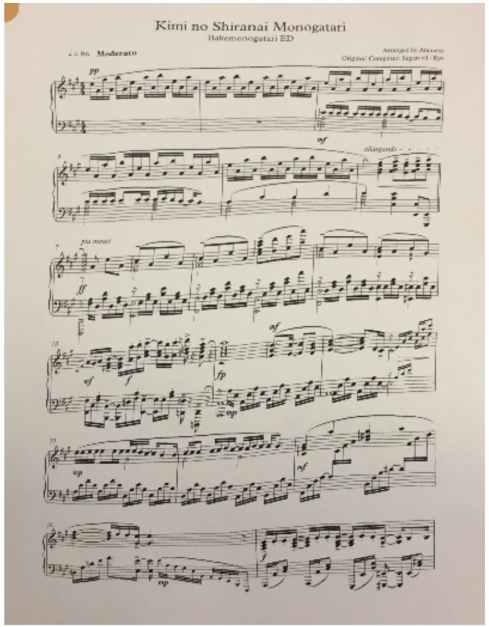


Fig 1: A pictorial overview of the functionality of our model.

Previous works

- A lot of work has been done in the domain of document dewarping using classical Image Processing based methods [5]-[6] and Optimization algorithms [7].
- Previous methods for Document Dewarping can be broadly classified as :
  - Use of stereo cameras to identify folds and deformation in documents: *Not easy to reproduce.*
  - Use of Image processing techniques to identify folds and then using homographic transforms to dewarp document: *Fails in cases of too many folds.*

However, we have seen relatively less use of Deep Learning algorithms in here.

Why not deep learning?

1. Shortage of large scale captured document images.
2. Scanned Documents cannot be used as ground truth--a simpler task needed.
3. Solved by recent works --DocUNet[1] and DewarpNet[2], which simulate warped documents and corresponding dense-grids as ground truths for the same.
4. A dense grid is a set of points that maps the coordinates from the original warped documents to coordinates in the dewarped image.
5. DocUNet and DewarpNet generate 100k data each, along with ground truths in various formats--aka 3d Coordinates, dense-grids (backward mappings), albedo maps and normal maps.

Novelty:

1

- Our network takes in 256x256 images as inputs to produce an unwarping grid which can be interpolated to reconstruct images at their original resolution.
- The parameters are learned efficiently and as such the model learns in just 8 percent of the dataset used in previous end-to-end methods.

2

- We propose a bifurcated U-Net as the secondary U-Net of our stacked U-Net system.
    - This helps in channel level segregation while predicting dense grid unwarps.

3

- A gated branch of the primary U-Net is also proposed.
    - This enables the secondary U-Net to recognize lines and boundaries in the warped document image.

Model architecture(a):

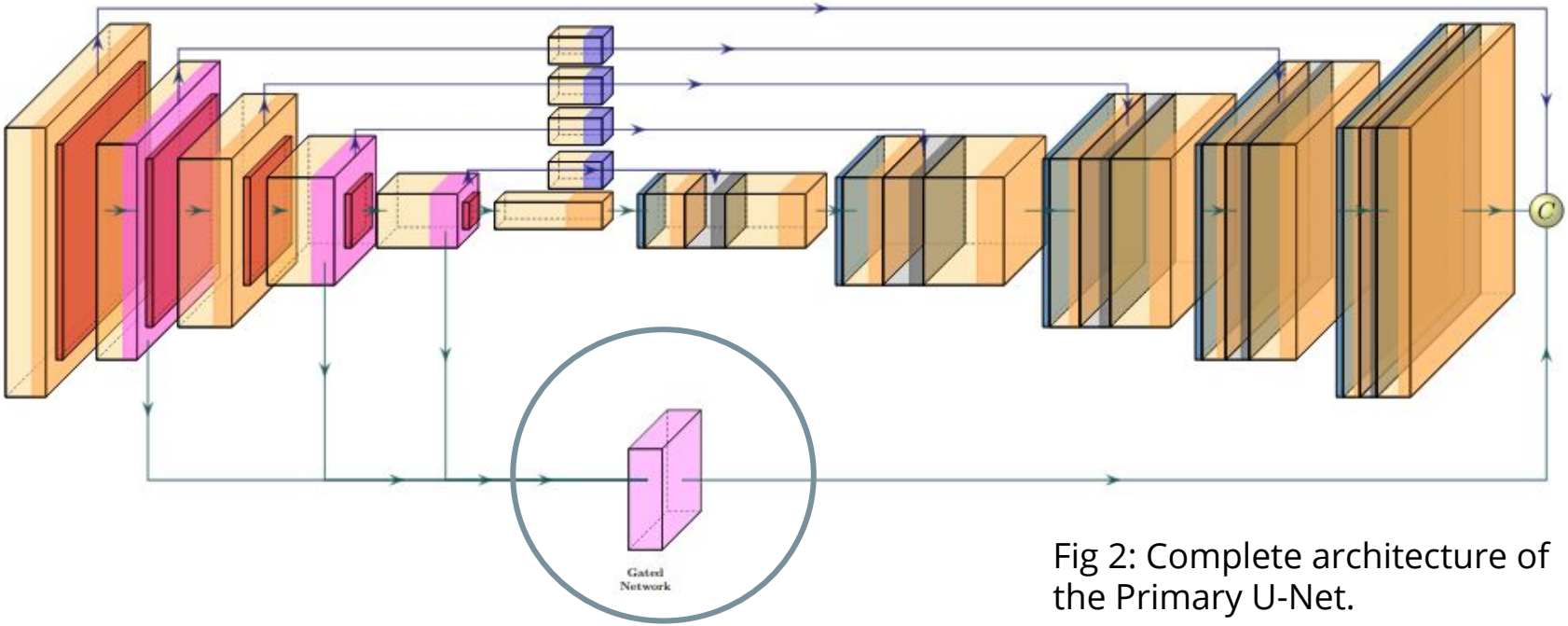


Fig 2: Complete architecture of the Primary U-Net.

Gated Network (GCN):

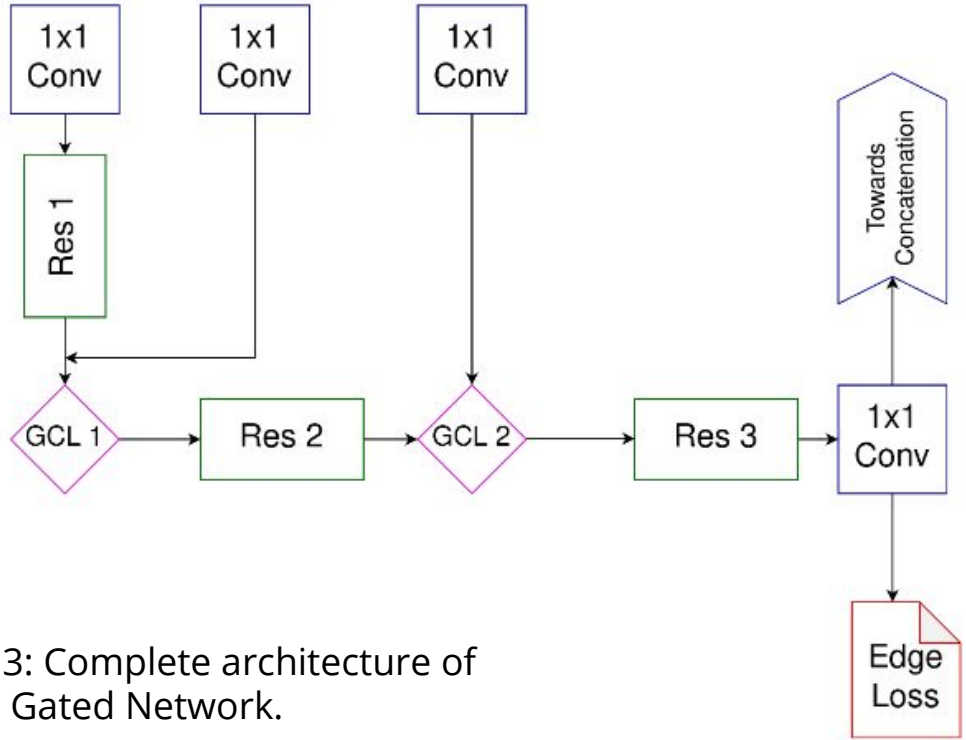


Fig 3: Complete architecture of the Gated Network.

Model architecture (b):

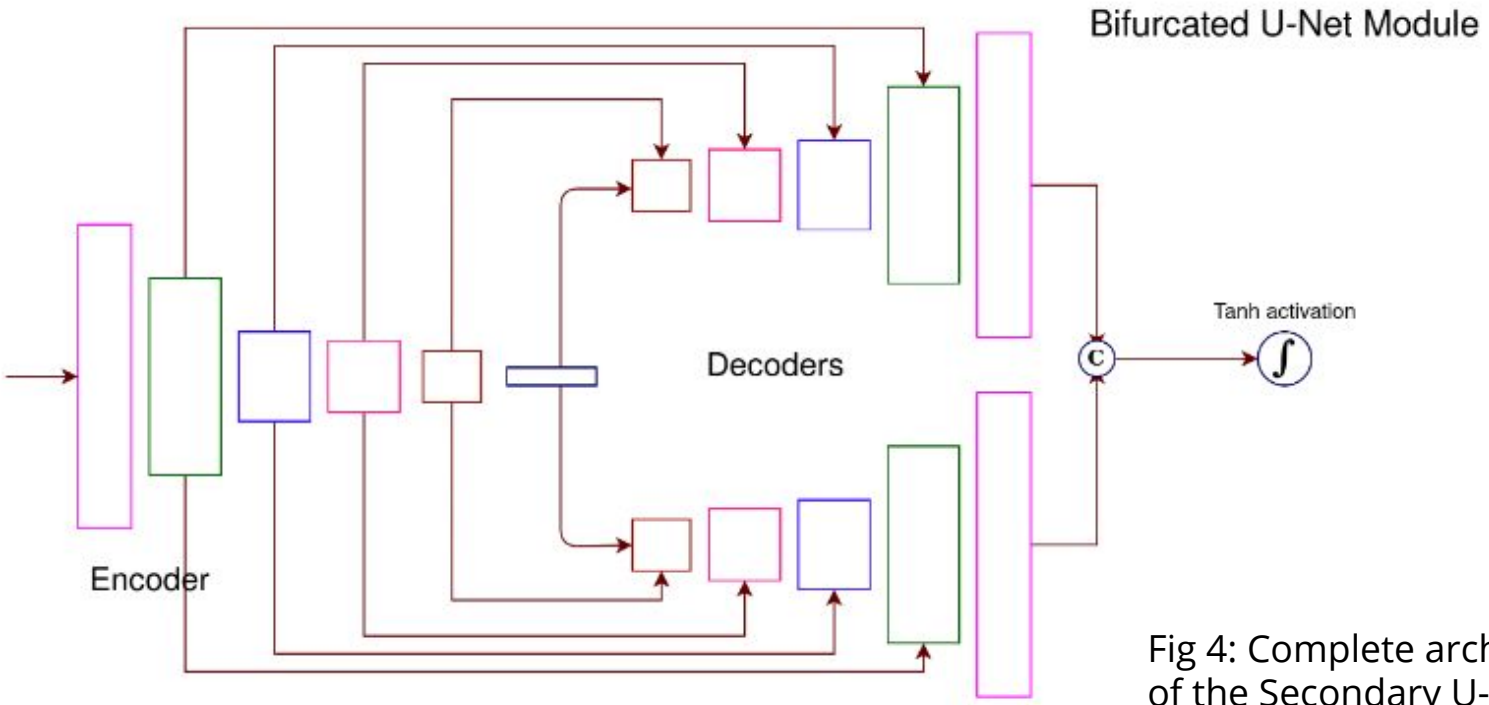


Fig 4: Complete architecture of the Secondary U-Net.

Why the split?

- The general CNN works by summing up computed data across all input channels for specific window sizes. Ultimately the number of channels in the output is the number of filters that the convolutional block contains.
- This summation results in merging of the data from multiple channels together into a single 2-dimensional vector and then using the merged data in the later stages.
- This is counter intuitive when we use this on dense-grid predictions as dense-grids channels don’t correlate and can’t be processed in this format.

Why the split?

- We came to the conclusion that using a single decoder in the final U-Net block would mean that although information is extracted in all blocks, it is merged together at each layer.
- Thus, only the last two convolutional filters would be responsible to decode or separate the grid values into their respective channels for the final output.
- To get round this issue, we came up with the usage of multiple decoder blocks for the single secondary U-Net encoder, so that channels in the dense grid output are developed separately.

Loss Function:

The loss function we use is a combination of :

- An Edge loss (BCE Loss), focussing on training the GCNs and the Primary U-Net in particular.
- A grid loss (MSE Loss) which trains the entire network.
- The summed up loss function is expressed as:

$$\mathcal{L} = \frac{1}{N}\sum_{i=0}^{N}(g_i - \hat{g}_i)^2 - \lambda.\frac{1}{N}\sum_{i=0}^{N} y_i.log(\hat{y}_i) + (1-y_i).log(1-\hat{y}_i)$$

We use the 0.9 as the value of lambda for all our experiments.

Results:

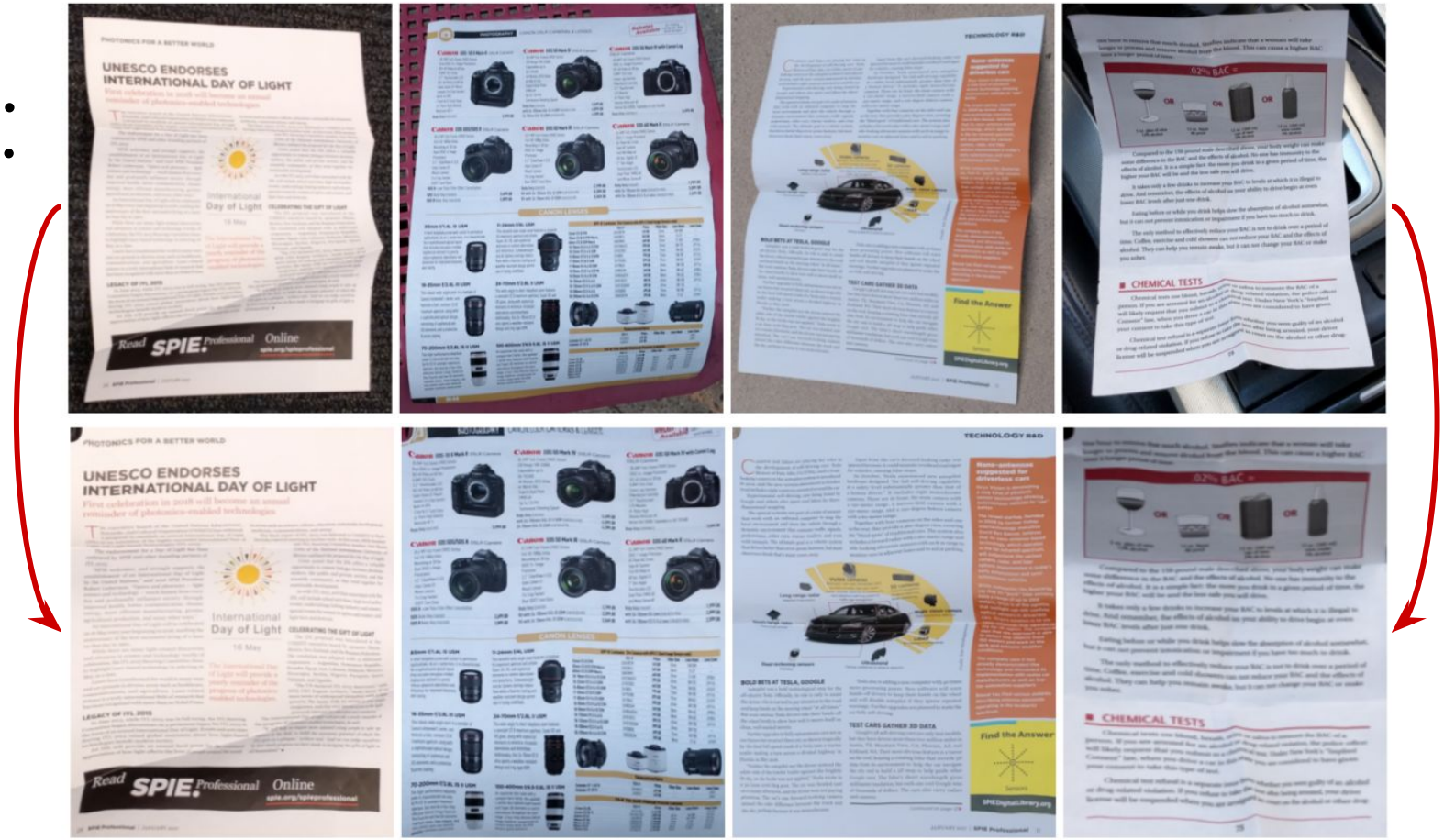


Fig. 4: Results on Benchmark introduced in [1].

SSIM & MS-SSIM

For comparing the quality of dewarp offered by our methods, after a minor post processing step, versus with those offered by previous methods, we make use of metrics like SSIM(structural similarity index) and MS-SSIM (multi-scale structural similarity index).

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

This measure is done between 2 windows x and y of common size NxN.

For MS-SSIM, A 5-level-pyramid is used where the weight for each level is set at **0.0448**, **0.2856**, **0.3001**, **0.2363** and **0.1333** [8].

SSIM along the levels:

Table 1:
Comparison with DewarpNet Results: Our Model can be seen to have a higher SSIM till the 3rd level. Level here refers to Gaussian Pyramid levels. The first level is the original image while the nth level has been down-sampled n-1 times.

Level	Our Method	DewarpNet [2]
Original Resolution	**0.548915**	**0.493146**
2	0.467136	0.433653
3	0.39162	0.387747
4	0.332977	0.369569
5	0.302610	0.464170
6	0.387984	0.575128
7	0.504144	0.607561
8	0.560574	0.586102
9	0.541162	0.546075

SSIM along the levels:

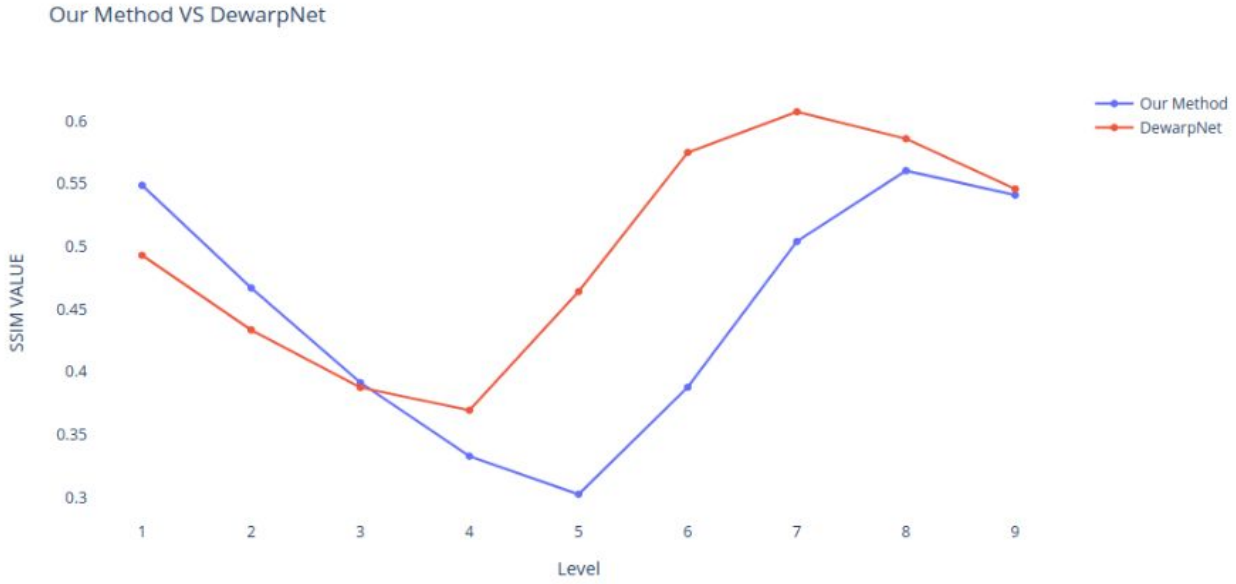


Fig 5: *Comparison with DewarpNet Results*: Figure illustrating the data from Table 1.

MS-SSIM and LD Comparisons:

Table 2: Comparison of MS-SSIM and LD with other methods in this domain.

Method	**MS-SSIM**	**LD**
Tian *et. al.* [3]	0.13*	33.69
DocUNet [1]	0.410*	14.08
Our Method	**0.415**	**13.2**
DewarpNet[2]	0.437	8.98

**Results obtained from research papers where data has been scaled.*

Future Work

- We find our methods lacking when applied on document images in the wild-- as we do not incorporate any specific kind of localisation procedure in our methods.
- A module for recognition and localisation of documents can help obtain better results when applied to document images.
- Additionally, we see that MS-SSIM as a metric does not provide as much attention to line level detail as it does to overall image structure, texture etc. The area dependency of MS-SSIM and LD also causes them to give highly varied results for the same distortion level in images of different areas.
- Thus, future work on an area independent standardized metrics is highly necessary for proper evaluation of results in this domain.

References

[1] X. B. J. W. D. S. Ke Ma, Zhixin Shu, “Docunet: Document image unwarping via a stacked u-net,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018.
[2] S. Das, K. Ma, Z. Shu, D. Samaras, and R. Shilkrot, “Dewarpnet:Single-image document unwarping with stacked 3d and 2d regression networks,” inProceedings of the IEEE International Conference on Computer Vision, 2019, pp. 131–140.
[3] Y. Tian and S. G. Narasimhan, “Rectification and 3d reconstruction of curved document images,” in CVPR 2011. IEEE, 2011, pp. 377–384.
[5] C. Wu and G. Agam, “Document image de-warping for text/graphics recognition,” in Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR).
[6] S. Lu and C. L. Tan, “Document flattening through grid modeling and regularization,” in18th International Conference on Pattern Recognition(ICPR’06), vol. 1.
[7] H. Ezaki, S. Uchida, A. Asano, and H. Sakoe, “Dewarping of documentimage by global optimization,” in Eighth International Conference on Document Analysis and Recognition (ICDAR’05).IEEE, 2005, pp.302–306.
[8] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, vol. 2. Ieee, 2003,pp. 1398–1402.

Thank you !