End-to-End Unsupervised Document Image Blind Denoising Supplementary Note

Mehrdad J Gangeh[1*] Marcin Plata[2*] Hamid R Motahari Nezhad[1] Nigel P Duffy[1]
[1]Ernst & Young (EY) LLP USA [2]EY GDS (CS) Poland Sp. z o.o.
{Mehrdad.J.Gangeh,Hamid.Motahari,Nigel.P.Duffy}@ey.com Marcin.Plata@gds.ey.com

1. Architecture of and training the proposed model

Table 1 provides a summary of the networks used in the proposed model (the details are provided in the main paper).

As explained in the main paper, we used in-house documents, including lease contracts, invoices, and tax forms to prepare the training dataset. The most common noise types on lease contracts, which are considered as unstructured documents, are S&P noise, blurred, or faded text, whereas tax forms (structured documents) and invoices (semi-structured documents) mostly contain watermarks.

The set of noisy and clean pages for the lease contracts are completely unpaired. As for the tax forms and invoices, extracting patches of 256×256 from the original watermarked pages resulted in only 10% patches with watermark (due to the fact that usually small part of the page is watermarked). Therefore, submitting these patches to the model did not train it adequately for watermark removal. To remedy this problem, we added watermarks similar to what are seen on actual tax forms or invoices, *i.e.*, with the same variations in text, font, size, orientation, transparency, and color to the grids of 4×2 of clean tax forms and invoices. This increased the number of watermarked patches by a percentage of more than 60%. A sample page with synthetically added watermark is shown in Figure 8a. The model was trained using the training dataset for 1,700,000 iterations.

2. Test sets used for quantitative evaluation

Table 2 provides additional information on the test sets used for the quantitative assessment of the proposed approach as reported in the main paper.

3. Additional results

3.1. Ablation study

In the main paper, we provided 10 consecutive values of a section of gating network g_H^* for the third convolutional layer of forward generator. Here, in Figure 1, we provide these values for eight remaining convolutional layers of forward generator. These values were calculated for two samples of all considered noise types, including S&P noise (blue), faded text (green), blurred text (yellow), and watermarked pages (red). These results are consistent with those displayed for layer three in Figure 2b of the main paper. As can be observed from these plots, there are strong similarities between the values generated by the gating networks for the same noise type, whereas they are different for different noise types. This demonstrates that the gating networks enable the forward generator to process an image in a different way depending on the containing noise type.

To further demonstrate the effectiveness of the gating networks, t-SNE [5] plots for all convolutional layers of forward generator are provided in Figure 2. The plots depict the distributions of gate outputs (256 features) reduced to two main components using t-SNE algorithm. The plots are obtained for 120 document pages containing one of the artifact types, *i.e.*, S&P noise, faded or blurred text, or watermarks (30 pages in each category). The plots for all layers show that the gates outputs are well separated for all four artifact types, which demonstrates the ability of the gating networks to separate the various noise types. From the plots, it can be observed that the least characteristic features are related to blurred pages, as they are sometimes overlapped with the features calculated on faded or watermarked images. On the other hand, it can also be observed that pages containing S&P noise make the most isolated class.

3.2. Qualitative results

We have provided more results on a few noisy document pages, including various artifacts, such as S&P noise, faded or blurred text, and watermarks. These results are provided in Figure 3 for a page from Tobacco800 dataset [6], Figures 4 and 5 for two sample pages from CDIP dataset [2], Figure 6 for a few samples from Kaggle dataset [1], and Figures 7, 8, and 9 for an instruction page of a tax form and two pages from a scientific paper with synthetically added watermarks, respectively. In order to demonstrate the ef-

[*]equal contribution

Table 1: The architectural details for the proposed model.

Networks	Details on the Components	Loss Function
Generators	ResNets (9 blocks)	GAN Loss + cycle-consistency loss
Discriminators	70×70 Patch-GANs	
Embedder (MoE)	CNN (7 layers, kernel: 3×3, batchnorm, ReLU)	Cross-entropy loss
Classifier (MoE)	Fully connected layer with softmax	
Gating Networks (MoE)	Fully connected (64×256), ReLU, 18 of these networks for the two generators)	ℓ_1 loss

Table 2: The details of the test datasets used for quantitative assessment of the proposed model.

Datasets	Dataset I	Dataset II	Dataset III				
	Scientific Papers	Tobacco800	Lease Contracts			Tax Forms	Invoices
Noise Types	Watermark	Various	S&P	Blurred	Faded	Watermark	Watermark
No. of Pages	100	100	60	100	60	40	40

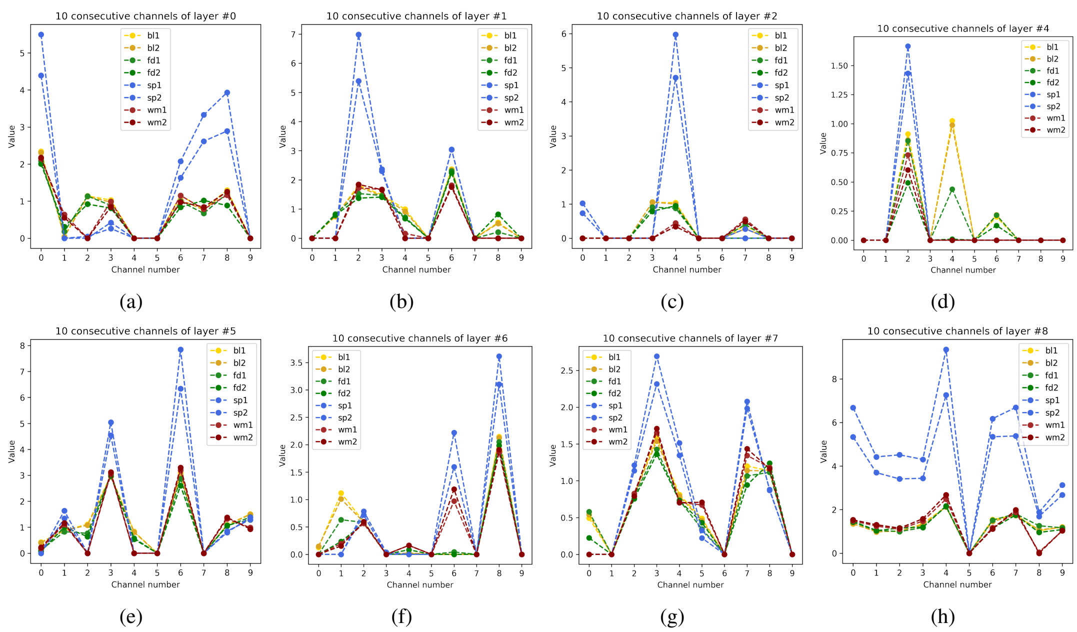


Figure 1: Ten consecutive values of a section of gating network g_H^* for all convolutional layers of forward generator (except layer three, for which the values are shown in Figure 2b of the main paper).

fectiveness of the proposed approach in image clean-up, we have compared our approach with the standard cycle-GAN [7] (without integrated deep MoE) trained on multiple noise types, including S&P noise, faded, and blurred pages in Figures 3 and 4. As can be observed from these two figures, the proposed model is much more effective in removing noise without distorting the texts on the pages. Furthermore, in order to demonstrate the improvement in the OCR after removing noise from a page, we have depicted the differences in OCR on part of a page before and after cleansing in Figure 5b. It can be observed that cleansing the page using the proposed approach is quite effective to improve the OCR performance and to generate correct OCR on the cleaned page. It is important to note that the proposed model has not been trained on any samples from Tobacco800, Kaggle, or CDIP datasets. As was explained in Section 1, the model has only been trained on our in-house documents, including lease contracts, tax forms, and invoices. Nonetheless, it produces excellent noise removal performance across these public datasets as demonstrated in the results depicted on this supplementary note.

For watermark removal problem, we have qualitatively

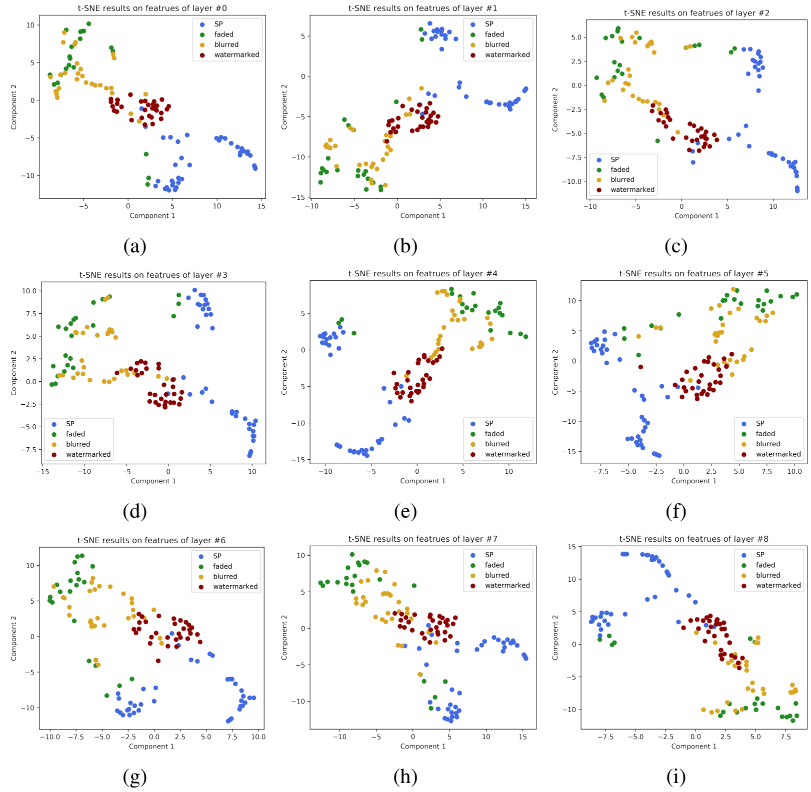


Figure 2: t-SNE [5] plots for all convolutional layers of forward generator provided for 120 document pages containing S&P noise, faded or blurred text, or watermarks.

compared the proposed approach with two supervised approaches, *i.e.*, REDNet [3] and DE-GAN [4] in Figures 7, 8, and 9. REDNet and DE-GAN have solely been trained using part of our training dataset containing paired watermarked/clean patches extracted from tax forms, whereas our proposed method has been trained on all noise types (S&P, faded, blurred, and watermarked). As can be observed in Figures 7c, 8c, and 9c, DE-GAN has difficulty to remove watermark from blank parts of the pages and this should be due to the additional loss function the authors have introduced to the model to preserve the text on pages (refer to Eq. (2) and corresponding explanations in [4]). Although our proposed model is unsupervised and has been trained for several noise types, it is as effective as REDNet (a supervised approach solely trained for watermark removal) in removing watermark from pages.

References

[1] Dheeru Dua and Casey Graff. UCI machine learning repository. Irvine, CA: University of California, School of Information and Computer Science, 2017. 1

[2] Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–995, 2015. 1

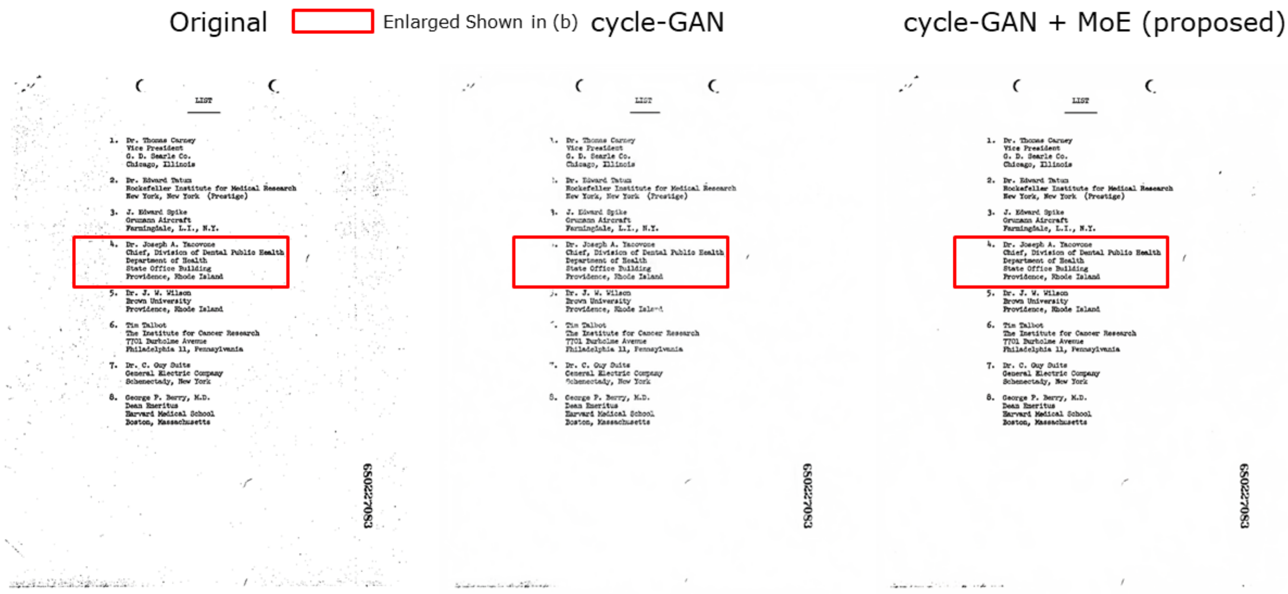
[3] Xiao-Jiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 2802–2810. Curran Associates, Inc., 2016. 3, 8, 9, 10

[4] Mohamed A. Souibgui and Yousri Kessentini. DE-GAN: A conditional generative adversarial network for document enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 3, 8, 9, 10

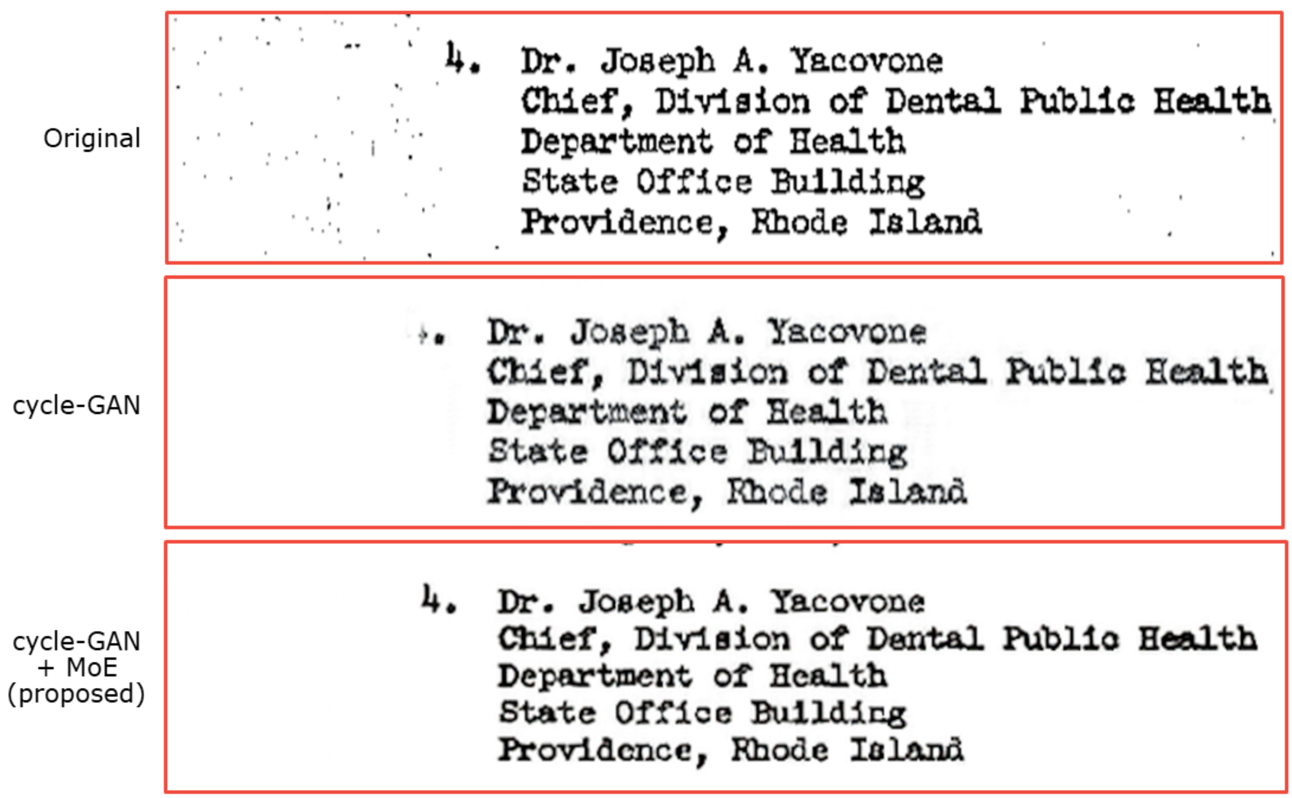
[5] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. 1, 3

[6] Guangyu Zhu, Yefeng Zheng, David Doermann, and Stefan Jaeger. Multi-scale structural saliency for signature detection. In *Proc. IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007. 1

[7] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017. 2, 4, 5

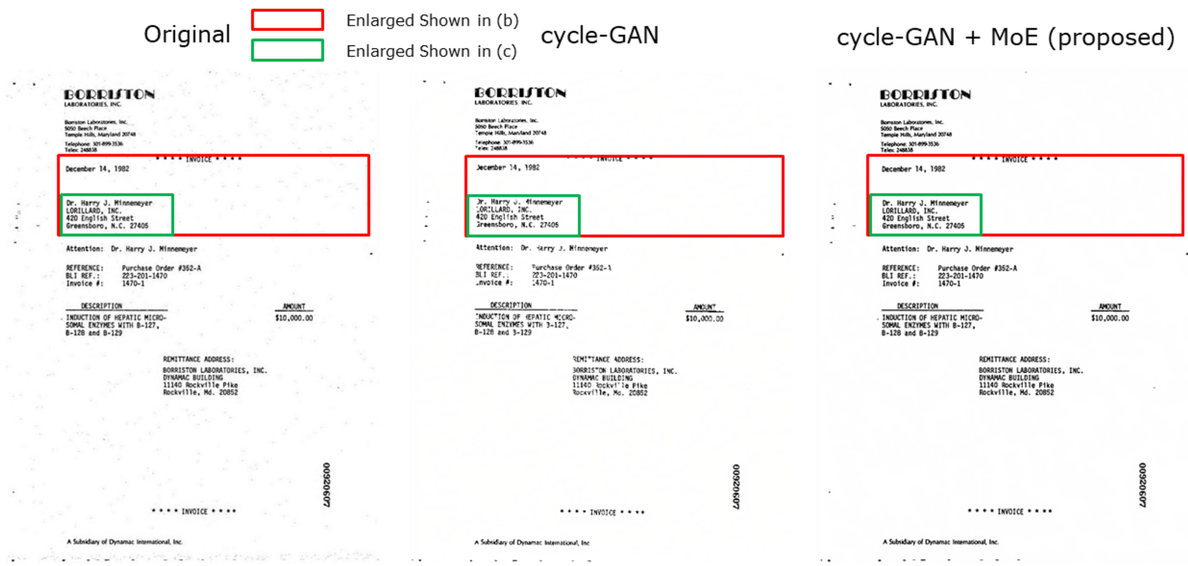


(a)

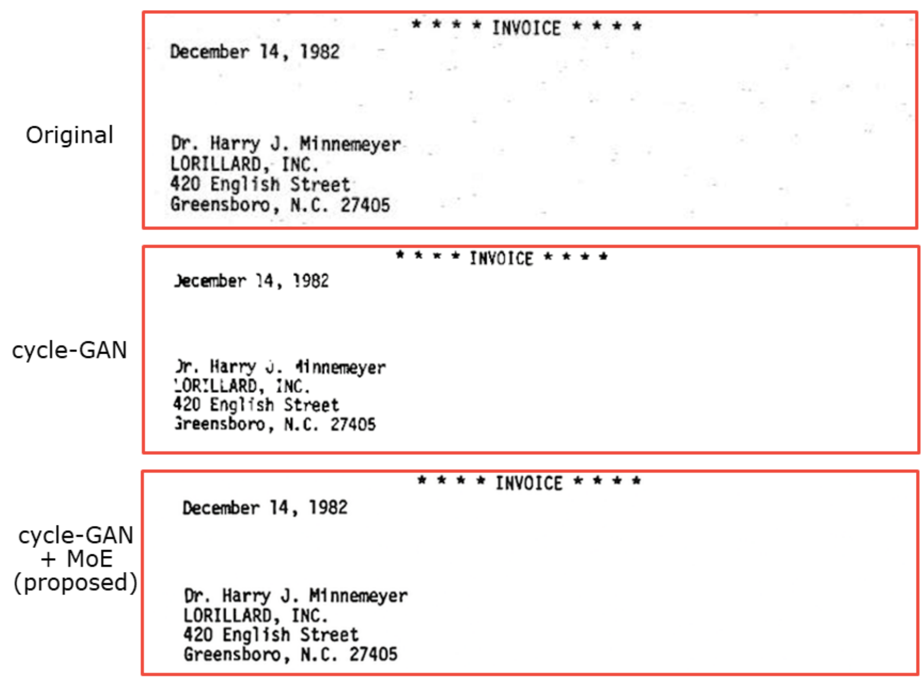


(b)

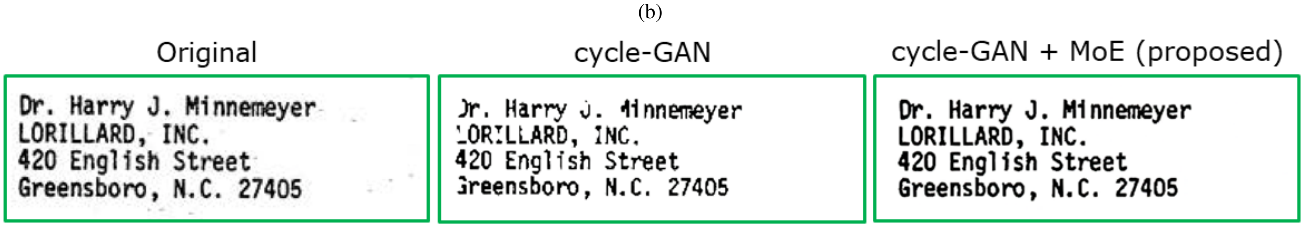
Figure 3: Qualitative results on a sample page from Tobacco800 Dataset (a) the whole page, (b) part of the page zoomed in (red box in (a)). The cleaned pages are compared between a standard cycle-GAN [7] and the proposed approach.



(a)

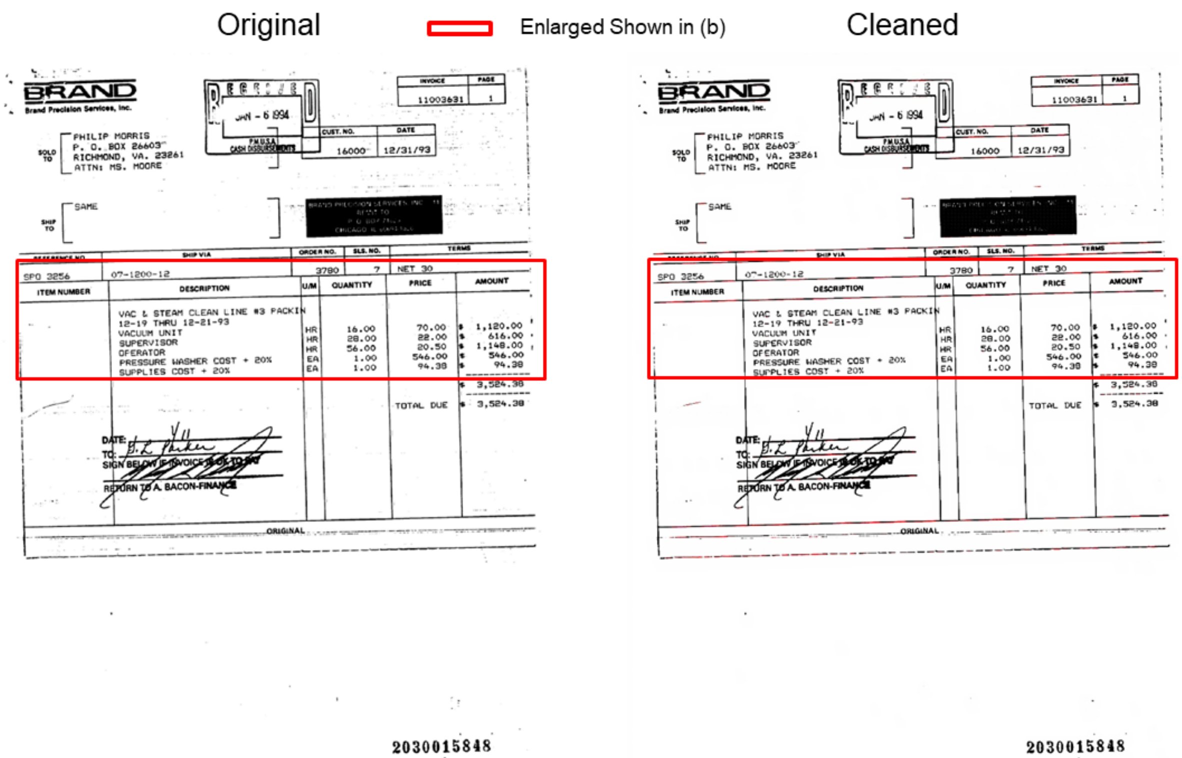


(b)

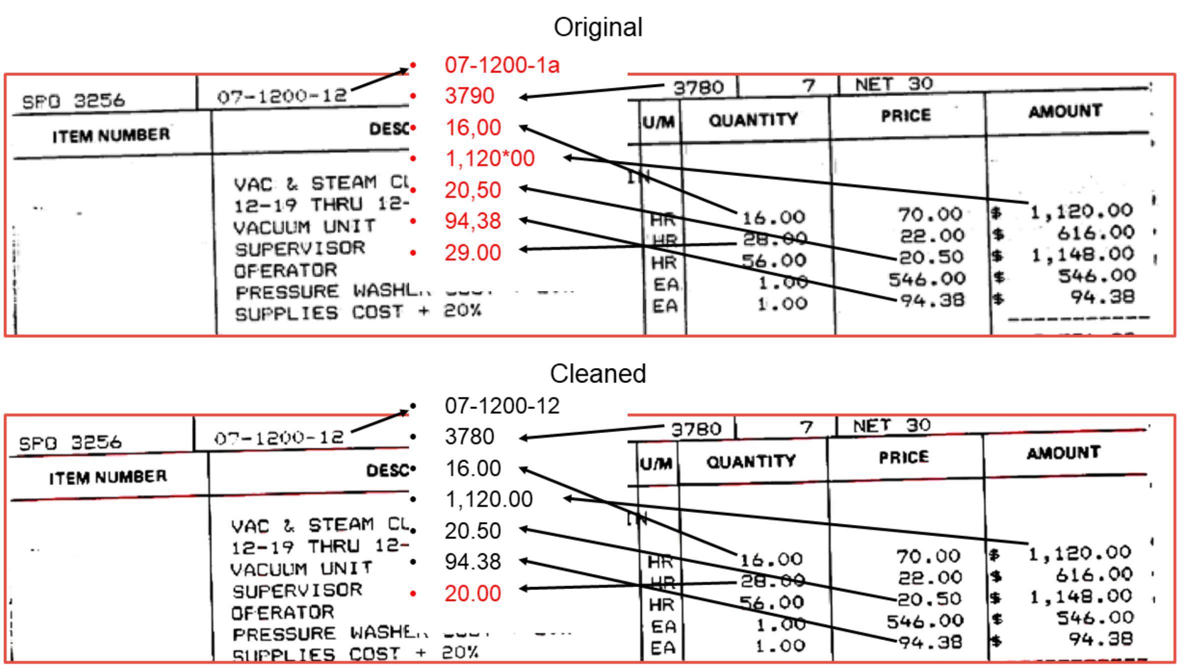


(c)

Figure 4: Qualitative results on a sample page from CDIP Dataset (a) the whole page, (b) part of the page zoomed in (red box in (a)), (c) part of the page zoomed in (green box in (a)). The cleaned pages are compared between a standard cycle-GAN [7] and the proposed approach.



(a)



(b)

Figure 5: Qualitative results on a sample page from CDIP Dataset (a) the whole page, (b) part of the page zoomed in (red box in (a)), the words that result in different OCR on the original and cleaned pages are also displayed for comparison.

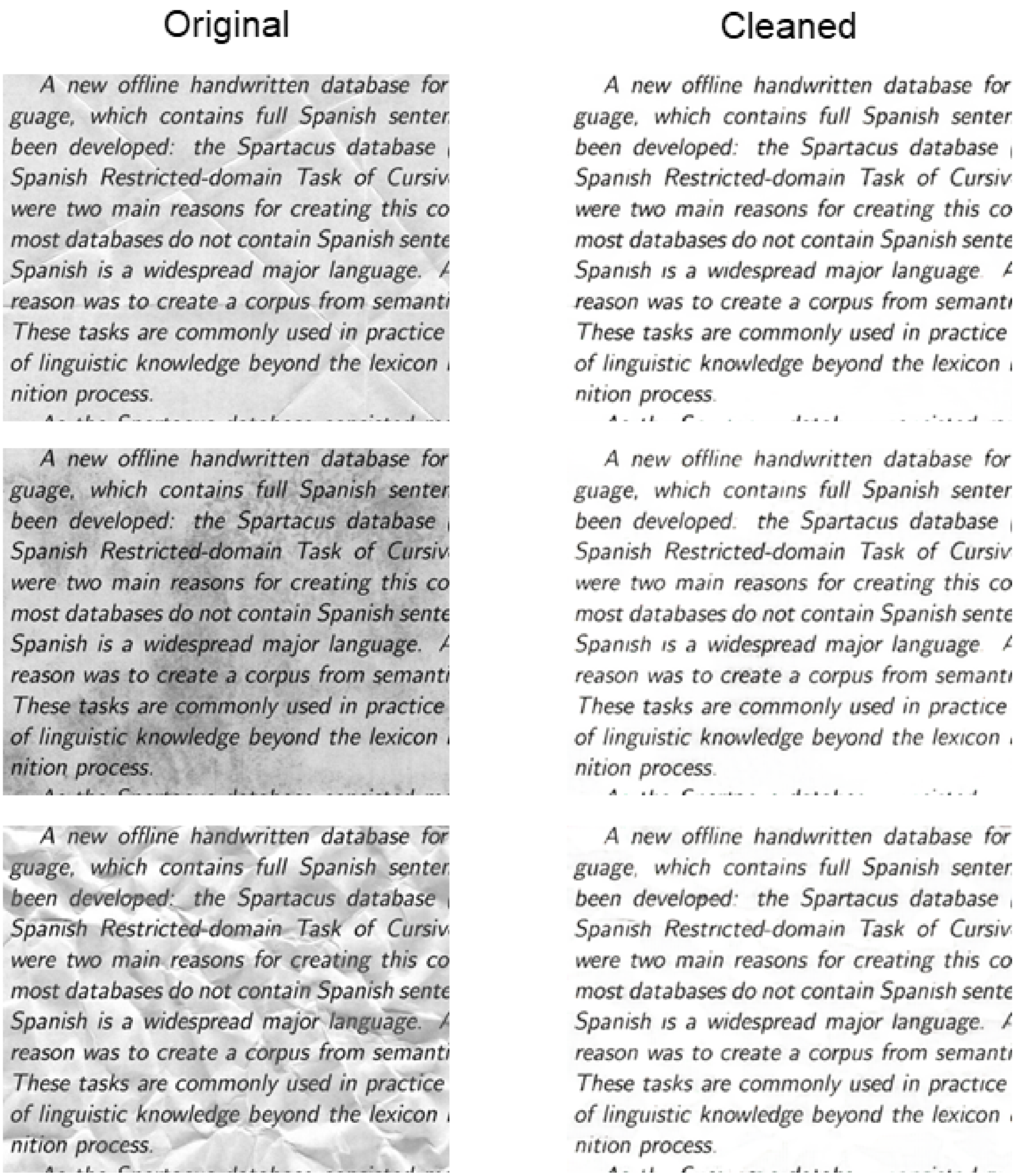


Figure 6: Qualitative results on a few samples from Kaggle Dataset.

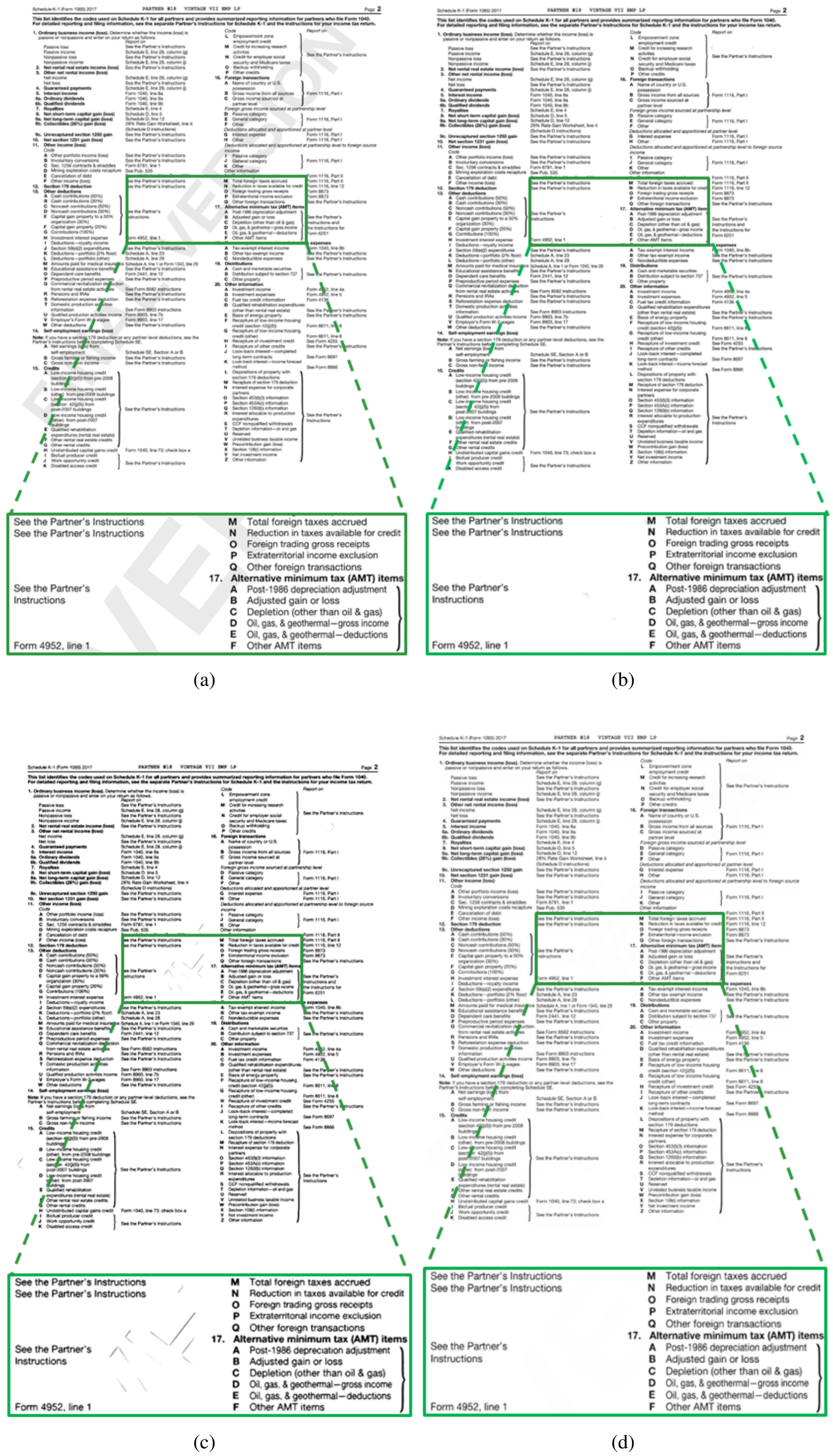


Figure 7: Qualitative results on a sample instruction page of a tax form: (a) watermarked page, and cleaned pages using: (b) RED-Net [3], (c) DE-GAN [4], and (d) proposed approach.

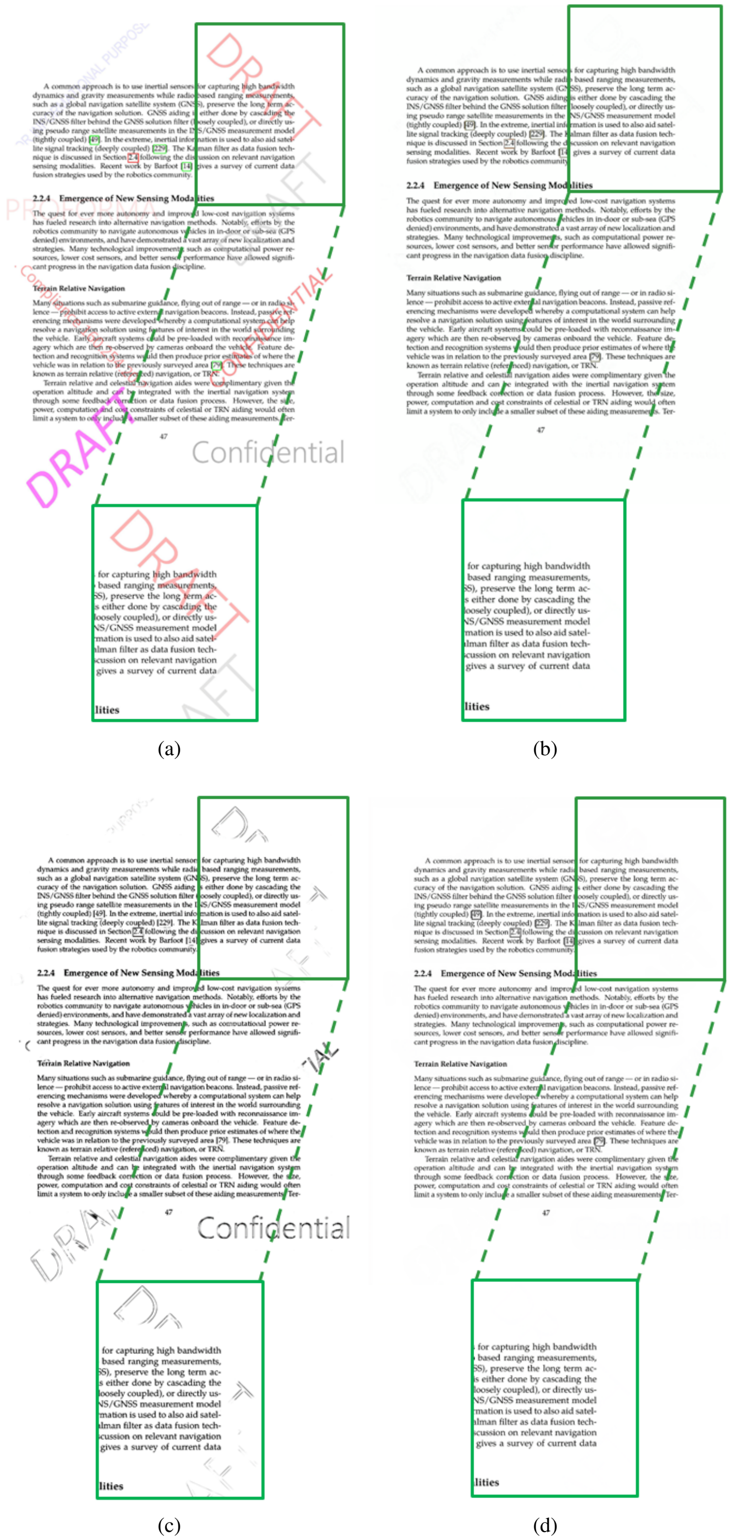
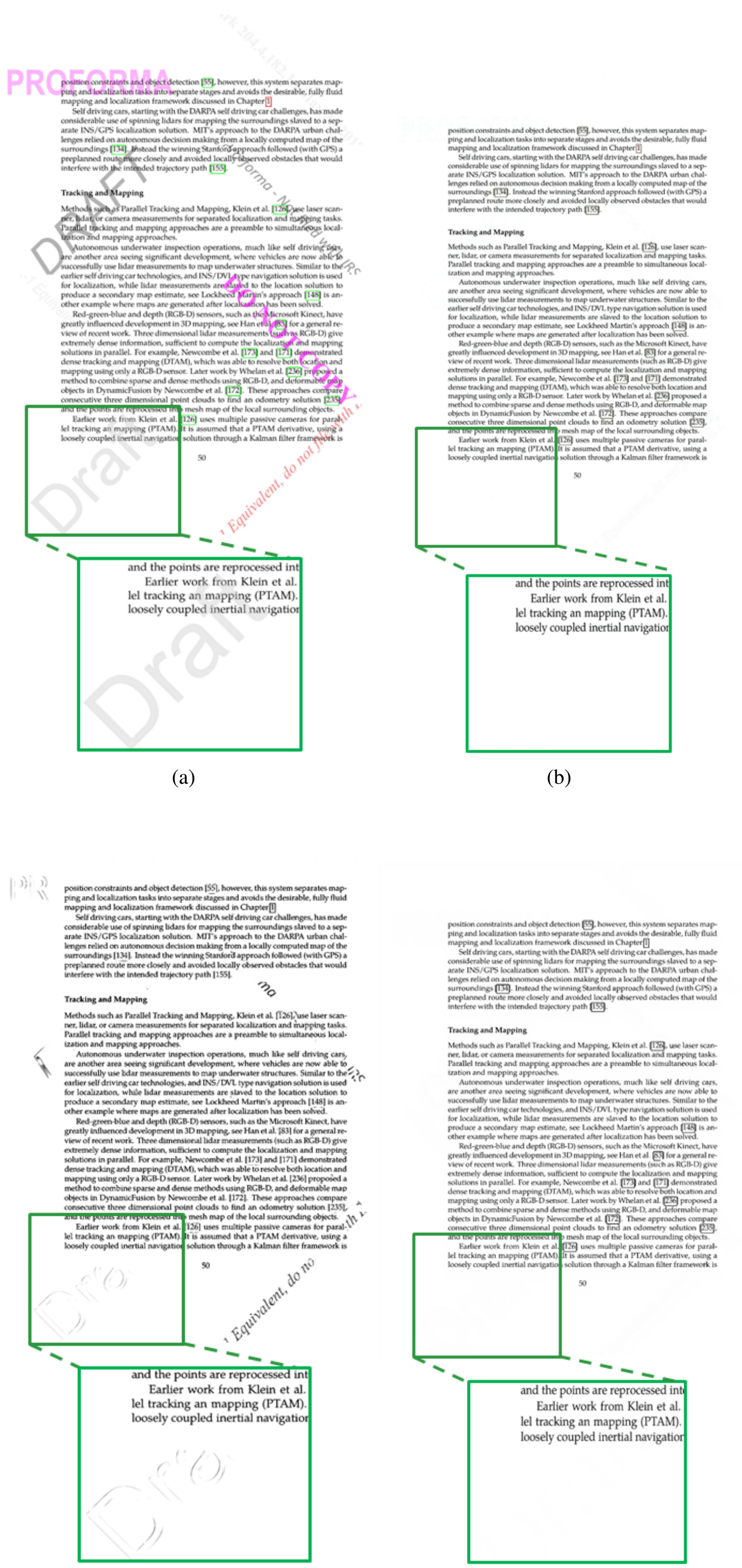


Figure 8: Qualitative results on a scientific paper with synthetically added watermarks: (a) watermarked page, and cleaned pages using: (b) RED-Net [3], (c) DE-GAN [4], and (d) proposed approach.



(a) (b) (c) (d)

Figure 9: Qualitative results on a scientific paper with synthetically added watermarks: (a) watermarked page, and cleaned pages using: (b) RED-Net [3], (c) DE-GAN [4], and (d) proposed approach.