Detection Masking for Improved OCR on Noisy Documents

Daniel Rotman*
danieln@il.ibm.com
IBM Research - Haifa
Israel

Ophir Azulai*
ophir@il.ibm.com
IBM Research - Haifa
Israel

Inbar Shapira
inbar_shapira@il.ibm.com
IBM Research - Haifa
Israel

Yevgeny Burshtein
bursh@il.ibm.com
IBM Research - Haifa
Israel

Udi Barzelay
udib@il.ibm.com
IBM Research - Haifa
Israel

ABSTRACT

Optical Character Recognition (OCR), the task of extracting textual information from scanned documents is a vital and broadly used technology for digitizing and indexing physical documents. Existing technologies perform well for clean documents, but when the document is visually degraded, or when there are non-textual elements, OCR quality can be greatly impacted, specifically due to erroneous detections. In this paper we present an improved detection network with a masking system to improve the quality of OCR performed on documents. By filtering non-textual elements from the image we can utilize document-level OCR to incorporate contextual information to improve OCR results. We perform a unified evaluation on a publicly available dataset demonstrating the usefulness and broad applicability of our method. Additionally, we present and make publicly available our synthetic dataset with a unique hard-negative component specifically tuned to improve detection results, and evaluate the benefits that can be gained from its usage.

CCS CONCEPTS

• **Information systems → Document structure**; • **Applied computing → Document analysis**; *Optical character recognition*; • **Computing methodologies → Object detection**.

KEYWORDS

Masking, Text Detection, Document Analysis

1 INTRODUCTION

Detecting and recognizing words and characters in images is a cornerstone technology for information extraction in the visual domain [30]. The difficulty of the task can often be divided into two categories: Optical Character Recognition (OCR) is often incorporated when the image is a digitized (scanned) document consisting mostly of aligned text in standard fonts displayed on uniform backgrounds [37]. For text appearing in natural images, Natural Scene Text recognition (NST) is often used, which incorporates advanced methods to overcome non-uniform backgrounds, non-standard fonts, and words appearing at odd angles or which have undergone spatial transformations [31]. Often the benefits of using NST are less pronounced when dealing with scanned documents, and the extra computation power, especially when dealing with the large amounts of words appearing in a standard document, makes it less of a viable option. Therefore in this work we focus mostly on OCR solutions for extracting text from documents.

One of the most broadly used solutions for OCR today is Tesseract [34]. Tesseract is considered a commodity and the go-to solution when the given task is text extraction from documents. The lightweight framework, multi-language support, ease of use, and open-source code provide an extremely useful resource. However, Tesseract exhibits degraded results on documents exhibiting non-ideal conditions [35]. Specifically, Tesseract tends toward false detections when there are noise artifacts or non-textual elements in the document such as logos, figures, and graphical elements.

In this work we propose a system to improve OCR results on degraded documents. Specifically, we create a pipeline which can be easily and readily applied to improve a standard OCR platform such as Tesseract. The core concept is to apply a pre-processing step with a designated detection network to perform a masking operation before processing the document with Tesseract (see Figure 1).

It is important to note that the main purpose of using this pipeline is to utilize the OCR's ability to perform better recognition when working at the document level. When using the same detector and applying OCR on every detection separately, the results are not as satisfactory due to the fact that contextual data could not be leveraged. Indeed, even if the detection network within Tesseract's pipeline could be replaced with the detector used for masking, results would likely not be better, as leveraging the contextual data was learned from clean documents. When the input is a noisy document, the artifacts given in the image will hamper leveraging the contextual element even when given a noise-free detector due to the appearance of the text's local visual surroundings.

To train a document text detector, we present a deep learning U-Net architecture [32] trained on our presented dataset. The dataset is synthesized with a variety of noise and difficult backgrounds as well as novel hard-negative samples to promote training of robust text detectors. We make the dataset publicly available as a standalone pre-generated archive of 100k documents.

We perform a series of evaluations demonstrating the usefulness of the dataset, our network, and the masking approach. Of note is the fact we demonstrate the performance on SROIE [13], a publicly available dataset of scanned receipts which accurately represents documents under difficult conditions.

Our contributions are as follows:

- We present our new masking formulation for improving OCR results on noisy documents.
- We present a U-Net architecture and training methodology, and demonstrate its usefulness for text detection in documents.
- We propose a new data synthesis approach with a novel hard-negative component, and make it publicly available [1]

*Both authors contributed equally to this research

[1] https://github.com/ophirazulai/SyntheticNoisyDocsDataset

Figure 1: A diagram of our system. The image is analyzed by the U-Net text detector and then undergoes a masking operation to eliminate non-textual artifacts. The clean document fed into the OCR technology results in superior text detection and recognition.

as a pre-generated dataset for the problem of text detection, and show the benefit of training given this dataset.

2 PREVIOUS WORK

2.1 OCR

For the task of OCR, the most prominent and wide-spread solution is Tesseract [34]. Originally, Tesseract was released as open-source in 2005 using mainly classic computer vision techniques including edge connected components analysis, blob filtering, quadratic spline fit, and recognition using topological features, and polygonal approximation. Starting from version 4, Tesseract employs LSTMs [8, 11] and the CTC loss [7] which are the current state-of-the-art approaches for text recognition.

Tesseract is considered the state of the art for OCR with regard to a commodity which is broadly applicable and easy to operate. The fact that it is so broadly used makes creating techniques to improve Tesseract's results doubly useful. For this reason we present our masking technique as a general pre-processing step which can easily improve Tesseract results and can be readily utilized.

However, when time or resource constraints are not an issue, advanced methods for extracting text from images exist. Natural Scene Text recognition (NST) networks have risen recently in popularity [31], but these approaches are outside of the scope of this work which focuses on light-weight text extraction from documents.

In this work we focus mainly on using the proposed masking approach before applying Tesseract as an OCR engine. However, the same reasoning and methodology can apply to other OCR systems.

2.2 Detection

For detecting multiple objects from images, a very common approach is creating segmentation maps using a U-Net architecture [32]. The architecture is characterized by the convolution layers which condense the spatial element to a bottleneck, and then up-convolutions which return the semantic information to the original spatial dimensions. There is a variety of uses for the U-Net architecture and variations, including image segmentation (with a wide usage in medical imaging) [14, 15, 46], but also in other tasks such as saliency detection [10], or as GAN descriminators [33].

Many powerful text detectors are constructed with architectures to promote NST. EAST [45] features a contracting and expanding network similar to the U-Net, and performs regression on the quadrilaterals based on the feature which is also used to generate the score map. PAN [40] also adopts a contracting and expanding network, and adds explicit kernel learning to isolate and better separate close text in the pixel aggregation stage.

2.3 Masking

In this work we use masking of non-textual elements as a type of de-noising technique to enable Tesseract to utilize contextual information without considering noise and artifacts [6].

The term masking can sometimes refer to attention modules [12, 41], or as part of training transformer networks [22]. Since the methodology we adopt is to perform the masking as a pre-processing stage and to leave Tesseract as a black-box, the methods above are not equivalent to our masking action which does not integrate knowledge or share semantic information.

Straightforward de-noising methods try to model the noise and convert the document accordingly [24, 29, 35, 38]. These methods do not rely on the use of a text detector which can be seen as an advantage, but often are limited to specific types of noise and can degrade the quality, sharpness, and shape of the actual textual elements. In our approach, instead of trying to model the noise, we aim instead on learning the ability to isolate and detect the text regions despite the noise present.

Finally, a common way to improve OCR results without intervening in the detection and recognition process is through post-OCR error correction [4, 16]. These steps often leverage language models and information, and also represent one of the types of context which Tesseract uses to improve OCR quality when performing on the document-level. Typically the types of linguistic and context errors here do not overlap with the ones our masking approach tries to solve such as noise and non-text artifacts, therefore this domain is beyond the scope of our work.

2.4 Document Datasets

The availability of datasets to train document OCR is limited.

FUNSD [17] consists of 199 documents with roughly 31k word annotations. The tasks and goals presented with the FUNSD dataset include mainly spatial layout analysis and form understanding. With quantity of this magnitude, this dataset can be useful for training and evaluating the tasks which require the component of semantic understanding. However, for the task of text detection it is necessary to have a much larger collection to encompass the

low-level variability which exists when attempting to isolate text shapes from non-ideal backgrounds.

SROIE [13] is a dataset consisting of scanned receipts for the tasks of OCR and key information extraction. Despite the large number of samples, the word count per document on these receipts is not large enough to promote training text detectors from scratch. We do however leverage this dataset for evaluation in Section 4.

Some additional datasets are Brno [20] which includes mostly spatial and lighting variations, quality assessment [21, 26] which concentrate on motion and focus blur, SmartDoc [2] which is comprised of videos and only 10 documents, and some others [1, 5, 27, 36]. However, none of these contain enough data to reliably train a robust text detector for documents.

The exception to the above is DDI-100 [44]. This dataset includes 7000 documents which then undergo a variety of transformations. Despite the strength of utilizing real documents, we evaluate and show in this work that the variability and distortions in the dataset are not diverse enough to train a powerful text detector which is truly robust to noise.

In contrast, NST datasets have risen greatly in popularity in the past years [3, 9, 18, 19, 25, 39, 42, 43]. However, the challenges presented in these datasets including irregular fonts and artistic text shapes and layouts, do not correctly represent the types of situations that a document text detector needs to learn to overcome.

3 METHOD

3.1 Dataset

We now present our synthetic automatically generated dataset for text detection in documents.

We use a python framework with the PIL library to synthesize text on images. Backgrounds are selected at various set probabilities with the options of white, natural image, and texture. For the latter, the textures are converted to grayscale and then a contour filter and a random dynamic range pixel value stretch is applied.

Text is synthesized with font size ranging from 9 to 100 pixels and font randomly selected 80% from 20-30 common fonts and 20% from a large assortment of unique fonts. The text to synthesize is selected randomly from a wikipedia content database [28], which includes a large corpus of words, numbers, domains, dates, phone numbers, URLs, and more. The ground truth heat-map is generated as character-level bounding boxes to avoid protruding letters causing the background around smaller letters to be labeled as text (see Figure 3).

Font-level noise is randomly added chosen from speckled dots, binarization, and random spatial distortions. Random small rotations are added to represent miss-alignment for scanned documents. As a final step, document-level noise is added randomly in the form of blur, compression, or downsampling. This step represents expected distortions which are likely to appear during scanning or photographing a document.

To promote powerful negative sample filtering, we present a novel hard-negative synthesis approach to create particularly difficult data with which to train the detection network (see Figure 2). Characters are generated and cropped into quarter-sized segments. A crop is augmented by a random rotation and scaling and re-rendered to represent contours which are character-like but not actual distinguishable letters.

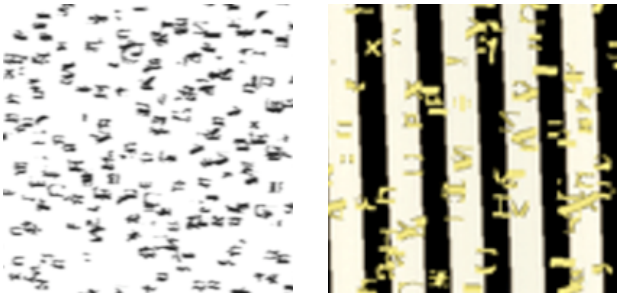


Figure 2: Examples of rendered hard-negatives. The detector learns to disregard character-like shapes and contours if they do not represent actual letters.

3.2 U-Net Network

We adopt a classic U-Net architecture for our text detection network. Four layers of convolution and then up-convolution are performed with 32, 64, 128, and 256 channels. In the up-convolution process skip-connections are employed by concatenating the output feature of the up-convolution with the feature of the regular convolution at the same level.

We use 100k synthesized documents 1024x1024 pixels, each consisting of a random number of synthesized tiles in each document (see Figure 3). The dice loss [23] is used with the ADAM optimizer and a learning rate of 1e-5.

3.3 Detection Masking

The output of the text detector is used to mask the given image. Areas which were not detected as containing text are blanked out and then the cleaned image is fed into Tesseract. Tesseract is run on the document-level, without providing the detections from the external text detector.

We note that a specific weakness of Tesseract is to sometimes identify entire paragraphs or lines as a single detection which

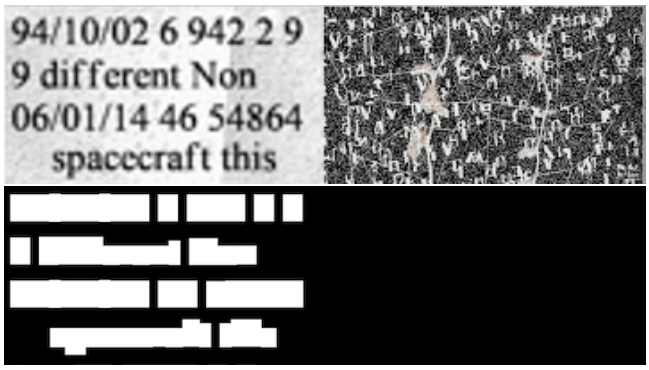
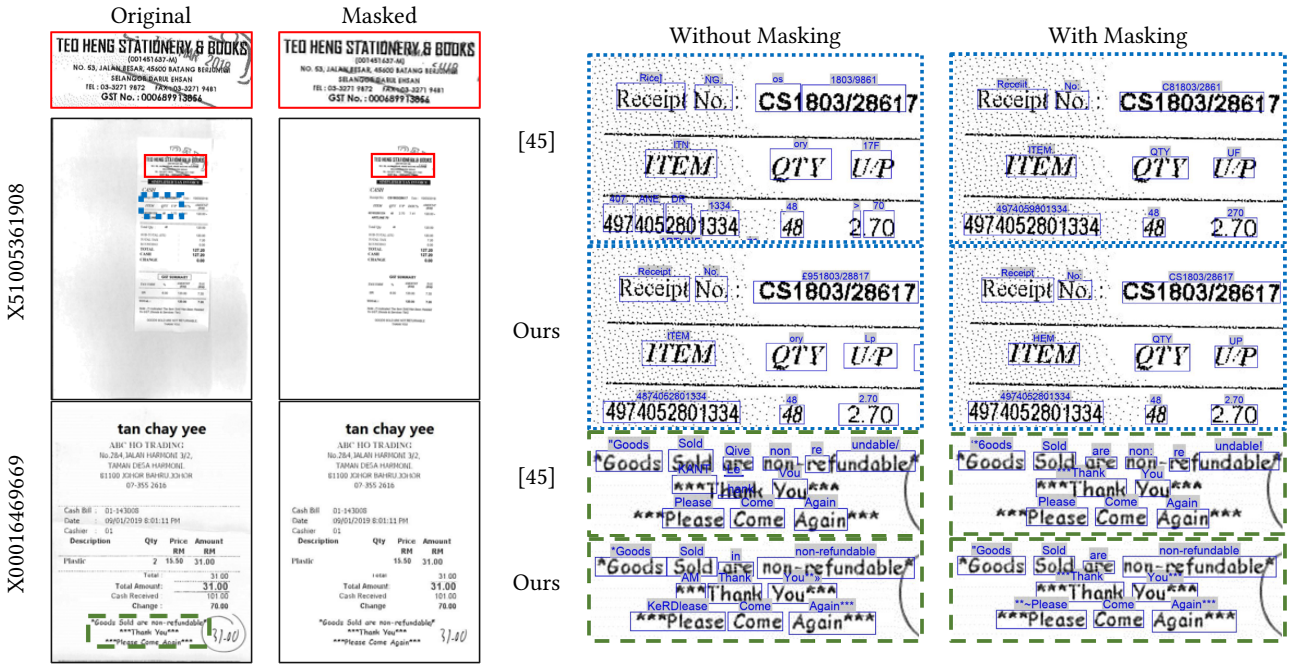


Figure 3: A miniature example of a generated document with only two synthesized tiles. Below is the matching generated binary ground-truth map for text localization.

Table 1: Visual results on two images from the SROIE dataset.



impedes recognition. Given the external text detection results, a post-processing procedure can be performed where these large detections are identified, merged with the text detector, and then resent to Tesseract for re-processing as individual detections.

4 EVALUATION

We evaluate our masking technique, detection network, and contribution of our dataset on the SROIE dataset [13]. We used all 1064 images, with annotated text bounding boxes and transcribed words.

In Table 2 we present the results of our evaluation. We measure detection using the F1-score of correctly detected bounding boxes using IOU 0.5. We measure recognition by the normalized edit distance for an entire document, where for True-Positives we calculate the edit distance as the Levenshtein distance, and for False-Positives and False-Negatives we set the edit distance as the length of the string. The normalized edit distance is the sum of edit-distances divided by the length of the text, and the Edit Score (ES) is 1 minus the edit distance. Case is rendered insensitive as the annotations of the SROIE dataset do not include upper or lower case.

Table 2: Results on SROIE dataset. Detection measured by F1-score, and recognition measured by average case-insensitive Edit Score (ES). 'Recognition' indicates using the method's detections for word-level recognition. 'Masked' indicates using our masking method and document-level recognition.

Method	Detection F1	Recognition ES	Recognition ES Masked
Tesseract	84	67	
PAN [40]	75	37	51
EAST [45]	84	48	59
U-Net - DDI [44]	82	56	62
U-Net - Ours	92	72	75

Tesseract results act as the baseline where the OCR is run at the document level. The consistent improvement which can be seen for all methods when using the masking technique shows the generality and usefulness of the approach. The improvement for U-Net between 'DDI' and 'Ours' shows the benefit of using out training dataset for the task of text detection.

In Table 1 we show some visual results from the SROIE dataset. On the left we show visually the output of the masking operation which results in a cleaner and often more eligible document. On the right we show close-up examples on the original document with the detection and recognition results visually embedded. 'Without Masking' represents using the detections and applying Tesseract on the word-level, while 'With Masking' represents using the masking technique and applying Tesseract on the cleaned output at the document-level.

5 CONCLUSIONS

In this work we presented a masking technique based on a designated text detector to improve document OCR. We introduced our new synthesized dataset with a novel hard-negative component designed to empower robust detection. Finally, through evaluation we showed the benefits of using the masking approach and of using the dataset to utilize OCR performance which utilizes contextual data on documents.

REFERENCES

[1] Konstantin Bulatov, Daniil Matalov, and Vladimir V Arlazarov. 2020. MIDV-2019: challenges of the modern mobile-based document OCR. In *Twelfth International Conference on Machine Vision (ICMV 2019)*, Vol. 11433. International Society for Optics and Photonics, 114332N.

[2] Jean-Christophe Burie, Joseph Chazalon, Mickaël Coustaty, Sébastien Eskenazi, Muhammad Muzzamil Luqman, Maroua Mehri, Nibal Nayef, Jean-Marc Ogier, Sophea Prum, and Marçal Rusiñol. 2015. ICDAR2015 competition on smartphone document capture and OCR (SmartDoc). In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1161–1165.

[3] Chee Kheng Ch'ng and Chee Seng Chan. 2017. Total-text: A comprehensive dataset for scene text detection and recognition. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 1. IEEE, 935–942.

[4] Deepayan Das, Jerin Philip, Minesh Mathew, and CV Jawahar. 2019. A cost efficient approach to correct OCR errors in large document collections. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 655–662.

[5] Iyad Abu Doush, Faisal AIKhateeb, and Anwaar Hamdi Gharibeh. 2018. Yarmouk arabic OCR dataset. In *2018 8th International Conference on Computer Science and Information Technology (CSIT)*. IEEE, 150–154.

[6] Filip Graliński, Tomasz Stanisławek, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. 2020. Kleister: A novel task for information extraction involving long documents with complex layout. *arXiv preprint arXiv:2003.02356* (2020).

[7] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*. 369–376.

[8] Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke, and Jürgen Schmidhuber. 2008. A novel connectionist system for unconstrained handwriting recognition. *IEEE transactions on pattern analysis and machine intelligence* 31, 5 (2008), 855–868.

[9] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. 2016. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2315–2324.

[10] Le Han, Xuelong Li, and Yongsheng Dong. 2019. Convolutional edge constraint-based U-net for salient object detection. *IEEE Access* 7 (2019), 48890–48900.

[11] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[12] Yunlong Huang, Zenghui Sun, Lianwen Jin, and Canjie Luo. 2020. EPAN: Effective parts attention network for scene text recognition. *Neurocomputing* 376 (2020), 202–213.

[13] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1516–1520.

[14] Nabil Ibtehaz and M Sohel Rahman. 2020. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Networks* 121 (2020), 74–87.

[15] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. 2018. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486* (2018).

[16] Adam Jatowt, Mickael Coustaty, Nhu-Van Nguyen, Antoine Doucet, et al. 2019. Deep statistical analysis of OCR errors for effective post-OCR processing. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 29–38.

[17] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, Vol. 2. IEEE, 1–6.

[18] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. 2015. ICDAR 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1156–1160.

[19] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. 2013. ICDAR 2013 robust reading competition. In *2013 12th International Conference on Document Analysis and Recognition*. IEEE, 1484–1493.

[20] Martin Kišš, Michal Hradiš, and Oldřich Kodym. 2019. Brno mobile OCR dataset. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1352–1357.

[21] Jayant Kumar, Peng Ye, and David Doermann. 2013. A dataset for quality assessment of camera captured document images. In *International Workshop on Camera-Based Document Analysis and Recognition*. Springer, 113–125.

[22] Bingcong Li, Xin Tang, Xianbiao Qi, Yihao Chen, and Rong Xiao. 2020. Hamming OCR: A Locality Sensitive Hashing Neural Network for Scene Text Recognition. *arXiv preprint arXiv:2009.10874* (2020).

[23] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*. IEEE, 565–571.

[24] Shashank Mujumdar, Nitin Gupta, Abhinav Jain, and Douglas Burdick. 2019. Simultaneous optimisation of image quality improvement and text content extraction from scanned documents. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1169–1174.

[25] Robert Nagy, Anders Dicker, and Klaus Meyer-Wegener. 2011. NEOCR: A configurable dataset for natural image text recognition. In *International Workshop on Camera-Based Document Analysis and Recognition*. Springer, 150–163.

[26] Nibal Nayef, Muhammad Muzzamil Luqman, Sophea Prum, Sebastien Eskenazi, Joseph Chazalon, and Jean-Marc Ogier. 2015. SmartDoc-QA: A dataset for quality assessment of smartphone captured document images-single and multiple distortions. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1231–1235.

[27] Wataru Ohyama, Masakazu Suzuki, and Seiichi Uchida. 2019. Detecting mathematical expressions in scientific document images using a u-net trained on a diverse dataset. *IEEE Access* 7 (2019), 144030–144042.

[28] Online. 2021. Database download. https://en.wikipedia.org/wiki/Wikipedia:Database_download

[29] C Patvardhan, AK Verma, and C Vasantha Lakshmi. 2012. Document image denoising and binarization using Curvelet transform for OCR applications. In *2012 Nirma University International Conference on Engineering (NUiCONE)*. IEEE, 1–6.

[30] Xujun Peng, Huaigu Cao, Srirangaraj Setlur, Venu Govindaraju, and Prem Natarajan. 2013. Multilingual OCR research and applications: an overview. In *Proceedings of the 4th International Workshop on Multilingual OCR*. 1–8.

[31] Zobeir Raisi, Mohamed A Naiel, Paul Fieguth, Steven Wardell, and John Zelek. 2020. Text Detection and Recognition in the Wild: A Review. *arXiv preprint arXiv:2006.04305* (2020).

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[33] Edgar Schonfeld, Bernt Schiele, and Anna Khoreva. 2020. A u-net based discriminator for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8207–8216.

[34] Ray Smith. 2007. An overview of the Tesseract OCR engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, Vol. 2. IEEE, 629–633.

[35] Dan Sporici, Elena Cușnir, and Costin-Anton Boiangiu. 2020. Improving the Accuracy of Tesseract 4.0 OCR Engine Using Convolution-Based Preprocessing. *Symmetry* 12, 5 (2020), 715.

[36] Uwe Springmann, Christian Reul, Stefanie Dipper, and Johannes Baiter. 2018. Ground Truth for training OCR engines on historical documents in German Fraktur and Early Modern Latin. *arXiv preprint arXiv:1809.05501* (2018).

[37] Nishant Subramani, Alexandre Matton, Malcolm Greaves, and Adrian Lam. 2020. A Survey of Deep Learning Approaches for OCR and Document Understanding. *arXiv preprint arXiv:2011.13534* (2020).

[38] Do Thanh-Ha, Salvatore Tabbone, and Oriol Ramos Terrades. 2013. Document noise removal using sparse representations over learned dictionary. In *Proceedings of the 2013 ACM symposium on Document engineering*. 161–168.

[39] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. 2016. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140* (2016).

[40] Wenhai Wang, Enze Xie, Xiaoge Song, Yuhang Zang, Wenjia Wang, Tong Lu, Gang Yu, and Chunhua Shen. 2019. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8440–8449.

[41] Zbigniew Wojna, Alexander N Gorban, Dar-Shyang Lee, Kevin Murphy, Qian Yu, Yeqing Li, and Julian Ibarz. 2017. Attention-based extraction of structured information from street view imagery. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 1. IEEE, 844–850.

[42] Sonia Yousfi, Sid-Ahmed Berrani, and Christophe Garcia. 2015. ALIF: A dataset for Arabic embedded text recognition in TV broadcast. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1221–1225.

[43] Ying Zhang, Lionel Gueguen, Ilya Zharkov, Peter Zhang, Keith Seifert, and Ben Kadlec. 2017. Uber-text: A large-scale dataset for optical character recognition from street-level imagery. In *SUNw: Scene Understanding Workshop-CVPR*, Vol. 2017.

[44] Ilia Zharikov, Philipp Nikitin, Ilia Vasiliev, and Vladimir Dokholyan. 2020. DDI-100: Dataset for Text Detection and Recognition. In *Proceedings of the 2020 4th International Symposium on Computer Science and Intelligent Control*. 1–5.

[45] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*.

5551–5560.
[46] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. 2018. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 3–11.