ShabbyPages: A Recipe for Repeatable, Synthetic, Modern Document Images

Anonymous ICDAR 2023 submission

No Institute Given

Abstract. The ShabbyPages document image dataset is produced using the Augraphy document image augmentation tool. The development of the computation pipeline used to generate the corpus is discussed, and the results presented. The final corpus contains over 6000 "born-digital" ground truth document images, sourced on the Web, with synthetically-noised counterparts ("shabby pages") that appear to have been printed and faxed, photocopied, or otherwise altered through physical processes. The release of this dataset and the recipe for its production attempts to address a growing need for labeled training document images for supervised learning tasks. The results of several experiments are discussed, in which the corpus trains performant convolutional denoisers, which remove real noise features with a high degree of human-perceptible fidelity, establishing baseline performance for a new ShabbyPages benchmark.

1 Introduction

Supervised learning requires a collection of training data and accompanying "labels", which in the graphical modality are clean images, free of noise or other degradations. ShabbyPages is a large dataset composed of 6202 clean/noisy pairs of images taken from 600 digital documents. Information about the set's contents and its relation to other datasets are explored in the next section. The noisy images were created from their clean origins by the application of a pipeline developed with the Augraphy library; the production of ShabbyPages is detailed in the third section, with all source code and input materials made openly available. The final dataset is used to train several denoising models which are compared directly and whose prediction results are compared together in the final section.

2 The Dataset

ShabbyPages joins a rapidly growing number of open-source datasets freely available online. Human activity today generates a tremendous amount of data, with many people choosing to publish theirs on websites like Kaggle [?], Hugging Face Hub [?], and many more. All of this data is arranged into categories: sets exist which contain text, images, audio, and data from other modalities, with each set often collected and designed for a purpose. Of the many image datasets available, the authors understand few to have been designed for groundtruthed supervised learning of document-specific tasks like binarization or denoising. In fact, we could find only one other set compiled for doing so with images of *modern* documents: NoisyOffice [1].

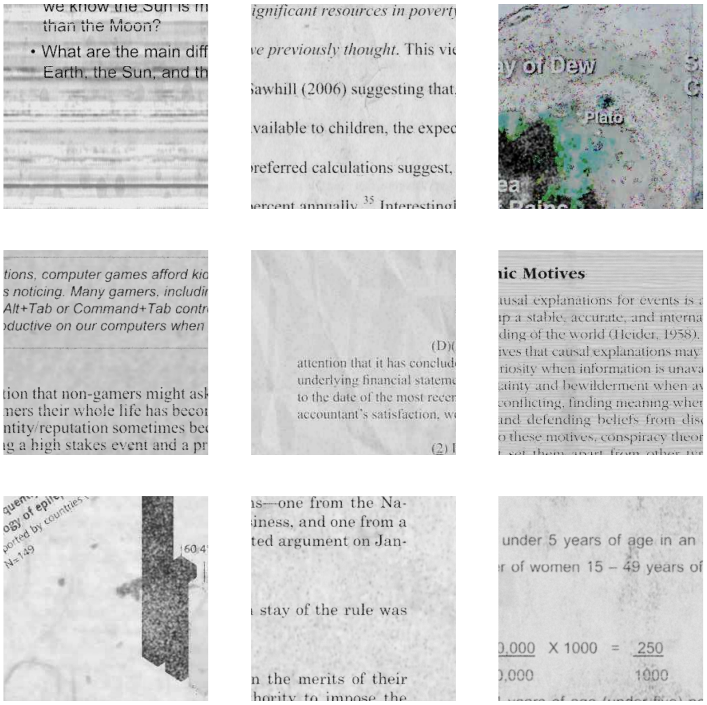
we know the Sun is m
than the Moon?
• What are the main diff
Earth, the Sun, and th
ignificant resources in povert
ve previously thought. This vi
Sawhill (2006) suggesting that
available to children, the expec
preferred calculations suggest,
percent annually.35 Interestingl
ay of Dew
Plato
tions, computer games afford kid
s noticing. Many gamers, includi
Alt+Tab or Command+Tab contr
oductive on our computers when
tion that non-gamers might ask
ners their whole life has beco
ntity/reputation sometimes be
g a high stakes event and a pr
(D)(
attention that it has conclud
underlying financial statem
to the date of the most rece
accountant's satisfaction, w
(2) I
nic Motives
ausal explanations for events is
p a stable, accurate, and interna
ding of the world (Heider, 1958).
ives that causal explanations may
riosity when information is unava
ainty and bewilderment when a
conflicting, finding meaning whe
and defending beliefs from dis
o these motives, conspiracy theo
ogy of epile
orted by countries
N=149
60.4
ns—one from the Na-
iness, and one from a
ted argument on Jan-
stay of the rule was
n the merits of their
hority to impose the
under 5 years of age in an
r of women 15 – 49 years of
0,000 X 1000 = 250
0,000 1000

Fig. 1. Sample patches from ShabbyPages.

2.1 Features

Modern artificial neural networks produce a latent representation of their training data, a space from which outputs are sampled. In deep networks, these representations live within multiple layers, each of which corresponds to a feature of the input data. Document images are a particularly interesting category, as the data within the documents each picture represents is itself frequently multi-modal; the document may contain an image and its description, a table with statistics, multiple regions of structured and formatted text, and even other documents, making their internal representations by neural networks often quite complicated.

Networks with a wider variety of latent representations frequently perform better when generalizing their learned functions to other tasks. Procuring a large-enough volume of robust-enough training data is paramount. Several sets of such images exist, many intended for processing by deep neural nets; Table 1 below contains a comparison of some of the key hyperproperties of the ShabbyPages set compared with other popular document image datasets.

Diversity. We use the diversity metric as defined in [6,7] to help us measure how diverse a dataset is. This metric is defined as

$$Diversity(X) = \frac{1}{|X|^2} \sum_{a \in X} \sum_{b \in X} \|f(a) - f(b)\|_2$$

where $f(\cdot)$ is an embedding function mapping the input image into an n-dimensional vector space. Here, we use CLIP's ViT model [8], which embeds each image into a 512-dimensional embedding space. The intuition behind the diversity metric is that datasets where images are highly similar will have lower diversity scores, and datasets where images share less visual similarity will have higher diversity scores. A dataset with low diversity might not be representative enough of the real world, and low diversity may often correlate with task easiness.

Table 1. Document Image Dataset Feature Comparison

Feature	ShabbyPages	NoisyOffice	RVL-CDIP	Tobacco800
Images	6202	72	400000	1290
Diversity	0.488	0.317		
Document-specific	Yes	Yes	Yes	Yes
Synthetic	Yes	Yes	No	No
Groundtruthed?	Yes	Yes	No	No
Font size	Multiple	16pt	Multiple	Multiple
Paper styles	Multiple	1	Multiple	Multiple
Multilingual	Yes	No	No	No
Documents contain images?	Yes	No	Yes	Yes
Pre-categorized	No	No	Yes	Yes

NoisyOffice The NoisyOffice database [1] contains a very low 4 base images, 2 types of font, 3 font sizes, and 4 types of noise, for a total of 72 unique images. The groundtruth images were produced in a standard word processor, with the simulated noisy images produced by first adding noise to a blank sheet of paper, storing that as a background image after digital scanning, then overlaying the foreground text onto the noisy background texture. An additional set of real NoisyOffice images was produced, where the foreground text was printed

Table 2. Summary of datasets for document denoising and document binarization tasks.

Dataset	***Dataset Size***	***Synthetic Noise***	***Ground-Truths***	***Diversity***	***Font size***	***Paper styles***	***Multilingual***	***Contains graphics***	***Pre-categorized***
ShabbyPages (ours)	6,202	✓	✓	0.488	Multi	Multi	✓	✓	✗
NoisyOffice [1]	72	✓	✓	0.317	16pt	1	✗	✗	✗
DocCreator									
DIBCO-9 [?]									
DIBCO-10 [?]									
DIBCO-11 [?]		✗	✓		multi	Multi	✓	✗	✗
DIBCO-12 [?]		✗							
DIBCO-13 [?]		✗							
H-DIBCO-14 [?]		✗							
H-DIBCO-16 [?]		✗							
DIBCO-17 [?]		✗							
H-DIBCO-18 [?]		✗							
DIBCO-19 [?]		✗							
Bickley Diary [?]									
LS-HDIB [?]		✓							

on blank white sheets, the results of which were physically augmented by the introduction of footprints, coffee stains, wrinkles, and folds.

ShabbyPages began with 600 documents sourced from the Internet from culturally-important sources. A manifest is included in the dataset which contains an ordered table of the filenames, the languages present in the document, the download URL for the file, and an English name or acronym for the source organization or person. These documents were split into their constituent pages, which were then passed through an Augraphy pipeline we tuned over a month or two. This tuning amounted to tweaking numerical parameters for each of the 24 different augmentations from the Augraphy library. All or none of these may have applied to each input image, and only one pipeline run was completed per image. Each augmentation has a high degree of internal variability due to several calls to a pseudorandom number generator on which several transforms depend. The conjunction of all these factors implies ShabbyPages is a sample from a massive space. Indeed, the cosine similarity computed over OpenAI's CLIP representations of each output image averages to 0.49, much higher than NoisyOffice's 0.31. Augraphy does not yet have augmentations for footprints or coffee stains, so the dark regions and edges of the stains and prints were simulated with features introduced by other augmentations, but Augraphy does contain the PaperFactory transformation which selects and crops a texture from a local directory of paper images. The result of a PaperFactory application is the foreground text “printed” onto the background texture, similar to NoisyOffice’s construction.

2.2 RVL-CDIP

The RVL-CDIP dataset [5] consists of 400,000 grayscale images divided into 16 document classes, each with a similar number of elements. The dataset is pre-segmented into 320,000 training, 40,000 test, and 40,000 validation images, facilitating immediate use in machine learning applications. The classes represented are the following:

letter, form, email, handwritten, advertisement, scientific report, scientific publication, specification, file folder, news article, budget, invoice, presentation, questionnaire, resume, memo.

The images exhibit poor fidelity with what we can imagine their originals to have been, contain substantial noise, and many were scanned at around 100 pixels per inch, quite a low resolution for documents. None of the images in the CDIP set come with groundtruths.

The collection of RVL-CDIP is an impressive feat in its own right: nearly half a million scanned documents with real noise, already categorized. Augraphy makes synthetic documents, intended to simulate real noise without the need to physically produce it. One of the augmentations used in the ShabbyPages set is BadPhotoCopy, which mimics low-fidelity noisy or damaged scans.

2.3 Tobacco800

Tobacco800, like RVL-CDIP, is a subset of the Complex Document Image Processing collection, containing images of documents originally internal to tobacco companies, and released under the Tobacco Master Settlement Agreement. The corpus is composed of 1290 images, gathered and scanned at different times and by different devices. Several of the images in this set are consecutively-numbered pages, increasing the set’s utility for certain image analysis tasks [3]. Document images in this set vary from from 150 to 300 DPI with dimensions between 1200 pixels wide and 2500 pixels tall on the low end and 1600 by 3200 on the high.

Every document in the ShabbyPages groundtruth set is a multi-page document. The released set was exported at 150 DPI, but the original documents are also distributed, so re-exporting at a higher resolution is trivial. The smallest image is 532 by 532, with the largest 3336 wide by 12157 tall. 6202 images are present in ShabbyPages, but this number could easily be increased by simply running another Augraphy pipeline to generate more.

2.4 ShabbyPages Compared

Supervised learning tasks require labeled training data, called the ground truth. Much of the world’s data does not come with labels, which instead must be created by humans in some fashion. In Table 1, we indicated that NoisyOffice was the only other groundtruthed dataset we examined. This isn't strictly true: there are projects to add some types of groundtruth to RVL-CDIP [4,9] and Tobacco800 [10,11], but the images in these sets were produced by real scans, without an accompanying original digital document. Conversely, NoisyOffice contains images of real scanned documents intended for validation, as well as their digital groundtruth and synthetic training image data.

ShabbyPages includes not only the groundtruth images, but also the original digital documents used to produce them, and the software for fully reproducing the ShabbyPages set, allowing anyone to quickly produce a new set in the Shabby family, optionally re-exporting groundtruths at a different resolution beforehand. The 6202 images in the ShabbyPages release set are really just the first sample from the space of these documents; we encourage building your own. NoisyOffice is the only other set we examined that contains its digital provenance, but the released database materials lack a means of reproducing the simulated printing method used to create the NoisyOffice training data. Augraphy's PaperFactory augmentation is such a tool. The other sets are not even theoretically extensible in this way.

2.5 Statistics

ShabbyPages contains documents from multiple classes, which contain many types of information.

Table 2 below contains some statistics for the dataset, in grayscale, at a standard resolution of 150 ppi.

Table 3. ShabbyPages dataset statistics.

Statistic	**Value**
Num. Images	6202
DPI	150
Max image size	—

2.6 Limitations

The ShabbyPages collection is a comprehensive entrant to the public datasphere, but with a relatively low starting document volume, makes some important omissions.

First, we arbitrarily selected 600 as the number of source documents. The number of final images of pages depends on this, and so is similarly arbitrary.

There are only a few languages represented, which reflects a number of biases, including the following:

- The native languages of the document-searchers
- The locations from which the document-searchers searched
- The search techniques used by each document-searcher

This work does not account for them, but these biases are significant determiners on the set of input documents. We hoped to overcome this somewhat by picking a suitably large number of input documents, and settled at 600.

3 Construction

This section describes the dataset generation methodology; code for all of this is available on GitHub.

3.1 Data Gathering

A team of workers searched the public internet for PDFs of many different kinds. 600 unique documents were retrieved, totaling 6202 pages. A similar process was completed to collect paper textures on which to "print" the documents: 300 unique textures were gathered, all of which are either in the public domain or carry CC-0, CC-BY, or CC-BY-SA licenses. Care was taken to retrieve freely-available and attributable documents; a CSV file containing document and paper texture links (as of time of download), with their sources, is available and distributed with the dataset. We also manually reviewed the gathered documents to verify that no personal identity information was present within the corpus.

3.2 PDF to PNG

The pdftoppm tool from the poppler-utils package was used to split the PDF documents into individual pages. Each page was converted to a PNG image at 150dpi, a common printing resolution. Because the majority of these documents were created with the standard US Letter dimensions, this resulted in the most common image dimension being 1275 pixels wide by 1650 pixels tall.

```
pdftoppm document_name.pdf document_name -png -r 150
```

3.3 Developing the Pipeline

The Augraphy library presents an easy-to-use API for constructing feature pipelines, which has been designed for interoperability with other augmentation tools, and within the broader data ecosystem. While Augraphy's default pipeline has what we believe to be quite realistic defaults, we wanted a broader range of features from this dataset than those the default pipeline could produce. To address this, we broke all of the parameters for every augmentation constructor out into separate variables, tweaking these and committing the new pipeline to GitHub. We created an automated daily build in GitHub Actions to render a random selection of ground-truth images with the updated pipeline, then our team met frequently to discuss the output and make adjustments to the pipeline.

3.4 Processing Augraphy on a multicore system

Execution time is dependent on which augmentations are executed at runtime; an Augraphy pipeline can take several seconds to process large images. The library is under active development with performance enhancements underway, but the time cost to generate large datasets sequentially is prohibitive when dealing with thousands of files, so we use a multi-process pooling technique to distribute the workload across many processor execution threads. For each process, we generate a new pipeline, run the pipeline object on an image, and save the output, using the code below:

```
import os
import cv2
from multiprocessing import Pool
from pathlib import Path

input_path = Path(”/path/to/input/images”)
filenames = [(input_path / name)
  for name
  in os.listdir(input_path)]

pool = Pool(os.cpu_count())

def run_pipeline(filename):
    image = cv2.imread(filename)
    # returns the current Shabby Pages pipeline
    pipeline = get_pipeline()
    data = pipeline.augment(image)
    shabby_image = data[”output”]
    cv2.imwrite(filename.parent / f”{filename.stem}—shabby{filename.suffix}”

pool.map(run_pipeline, filenames)
```

Processing all 6202 images took less than half an hour on the 64 cores of a Graviton3 c7g.16xlarge instance on AWS.

4 Experimentation with the ShabbyPages Set

Binarization and denoising are two important techniques for the removal of unwanted data from images. Both approaches can be achieved by supervised learning of relationships between features of the input and output data. In this section, we train and compare several instances of the NAFNet [2] architecture as denoisers and binarizers, and evaluate these models' predictions on both the NoisyOffice and ShabbyPages datasets. The model was “off-the-shelf”; besides rigging up the training inputs, we made no alterations to the

4.1 Motivation for Denoising

Digital computation has enabled humanity to produce ever-growing amounts of both data and reasons to process it. Doing so is easiest when the data is free of unexpected outliers, the signal has less jitter, or when our aesthetic sensibilities aren't injured.

4.2 Binarization

In this section, we discuss results obtained by training a binarizing denoiser to both restore augmented images to their ground truths and to classify pixels as foreground.

We first performed an ensemble binarization preprocessing step on the clean groundtruth images. For each image, we computed the Niblack, Sauvola, and Otsu thresholds, using these to binarize three copies of the input. The average over the resulting matrices was taken elementwise, producing the final binarized groundtruth.

One instance of the model was trained on the SimulatedNoisyOffice corpus, while the other was trained on a subset of ShabbyPages, in both cases using the relevant ensemble-binarized groundtruths. As a test, these models were used to denoise and binarize both the NoisyOfficeReal dataset and a testing subset taken from ShabbyPages which does not overlap the training subset. Common image processing metrics were applied to the predictions made by each model, comparing these to the (preprocessed) groundtruths. The averages of each metric over all such predictions was taken; these are presented in Table 2.

Table 4. Document image binarization performance of a NAFNet model trained and tested on ShabbyPages and NoisyOffice.

Training Set	Test Set	SSIM↑
ShabbyPages	NoisyOffice-real	0.947
NoisyOffice-sim	ShabbyPages	0.811

Interpretation As a baseline, we first evaluated the models against test sets associated with their training databases; the model trained on SimulatedNoisyOffice was used to enhance the RealNoisyOffice set, while the ShabbyPages-trained model cleaned a testing subset drawn from the full ShabbyPages set.

We considered three common image processing metrics: structural similarity index, root mean squared error, and peak signal-to-noise ratio.

Although the ShabbyPages-trained NAFNet predicted cleaned ShabbyPages images with a lower structural similarity score than the NoisyOffice-trained model predicted cleaned RealNoisyOffice images, the ShabbyPages model achieved superior performance on RMSE and PSNR. The ShabbyPages corpus contains a much higher feature diversity than NoisyOffice, so the lower SSIM score was not unexpected; local relationships between pixels were easier to encode for the NoisyOffice model, possibly because there are simply fewer relationships. The RMSE and PSNR results on the other hand indicate that the ShabbyPages-trained model learned to remove more noise with higher fidelity, and more accurately predicted foreground pixels than the NoisyOffice model. This discrepancy can partly be explained by the RealNoisyOffice corpus having been physically-noised as opposed to ShabbyPages' purely synthetic noise; small-scale features in the RealNoisyOffice testing data are not quite identical to the Simulated-NoisyOffice training inputs, while both the training and testing data taken from ShabbyPages was constructed by the same means. In the latter case, the *diversity* of the ShabbyPages data is much higher (TODO: add diversity score info), and this instance of the NAFNet may have been able to generalize over noise types better than the NoisyOffice-trained network with its lower training diversity.

The second stage of the binarization experiment was to cross-validate these models by predicting cleaned images using the other testing set as inputs. Here, the ShabbyPages-trained model achieved all-around better performance than the NoisyOffice-trained model, with a much higher structural similarity score (0.947 vs. 0.811), lower RMSE, and higher PSNR. This is a strong indicator that neural networks trained on ShabbyPages can outperform those trained on NoisyOffice when generalizing to other datasets.

Interestingly, the NoisyOffice-trained model RMSE and PSNR scores were lower when cleaning RealNoisyOffice than the Shabby-trained model: we were able to achieve a 3.5 decibel improvement in peak signal-to-noise ratio over the SimulatedNoisyOffice-trained model.

References

1. Castro-Bleda, M.J., España-Boquera, S., Pastor-Pellicer, J., Zamora-Martínez, F.: The NoisyOffice Database: A Corpus to Train Supervised Machine Learning Filters for Image Processing. The Computer Journal **63**(11), 1658–1667 (11 2019). https://doi.org/10.1093/comjnl/bxz098, https://doi.org/10.1093/comjnl/bxz098
2. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. arXiv preprint arXiv:2204.04676 (2022)
3. Doermann, D.: Tobacco 800 dataset, https://tc11.cvc.uab.es/datasets/Tobacco800_1
4. Goldmann, L.: Layout analysis groundtruth for the rvl-cdip dataset (2019). https://doi.org/10.5281/ZENODO.3257319, https://zenodo.org/record/3257319
5. Harley, A.W., Ufkes, A., Derpanis, K.G.: Evaluation of deep convolutional nets for document image classification and retrieval. In: International Conference on Document Analysis and Recognition (ICDAR) (2015)
6. Kang, Y., Zhang, Y., Kummerfeld, J.K., Tang, L., Mars, J.: Data collection for dialogue system: A startup perspective. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics:

Human Language Technologies, Volume 3 (Industry Papers) (2018), https://aclanthology.org/N18-3005

7. Larson, S., Mahendran, A., Lee, A., Kummerfeld, J.K., Hill, P., Laurenzano, M.A., Hauswald, J., Tang, L., Mars, J.: Outlier detection for improved data quality and diversity in dialog systems. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) (2019), https://aclanthology.org/N19-1051
8. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021), https://arxiv.org/pdf/2103.00020.pdf
9. Riba, P., Dutta, A., Goldmann, L., Fornés, A., Ramos, O., Lladós, J.: Table detection in invoice documents by graph neural networks. In: International Conference on Document Analysis and Recognition (ICDAR) (2019)
10. Zhu, G., Doermann, D.: Automatic document logo detection. In: 9th International Conference on Document Analysis and Recognition (ICDAR) (2007)
11. Zhu, G., Zheng, Y., Doermann, D., Jaeger, S.: Multi-scale structural saliency for signature detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2007)