Evaluating Deep Neural Networks for Faithful Restoration of Historical Documents

Filip Tuch*
Supervised by: Zuzana Černeková†

Department of Applied Informatics
Comenius University in Bratislava
Bratislava, Slovakia

Abstract

The digitisation of historical archives is frequently hindered by physical degradation of documents, requiring costly manual restoration or automated approaches based on deep learning. Although deep learning offers powerful tools for restoration, its application in this domain raises significant safety concerns about semantic fidelity, specifically the risk of 'hallucinating' false characters that alter the meaning of the document. This paper evaluates and compares the safety and efficacy of three distinct deep learning architectures: a Generative Adversarial Network (Pix2Pix) and two Convolutional Neural Networks (DnCNN and DRUNet) in the restoration of degraded text documents. We introduce a synthetic dataset generation pipeline that simulates various types of degradation commonly found in historical documents, along with the extraction of precise textual ground truth for accurate evaluation. Beyond traditional image quality metrics, we propose a safety assessment framework that distinguishes between safe errors (deletions or data loss) and unsafe errors (substitutions or insertions) in the context of text restoration. Our results demonstrate that while the generative Pix2Pix improves visual quality and general readability, it is unsafe for document restoration due to the frequent introduction of 'hallucinated' characters. On the contrary, DnCNN and DRUNet exhibit safer restoration capabilities by minimising both safe and unsafe errors, achieving superior results in both semantic safety and image quality metrics.

Keywords: document restoration, deep learning, hallucination, semantic fidelity, safety assessment, image quality assessment

1 Introduction

Preservation of cultural heritage is highly dependent on digitisation of archives, to ensure that manuscripts, records, and literary works remain accessible to future generations. However, many of these documents suffer from physical degradation over time [1, 2], including fading ink, stains, tears, mould, and foxing, as well as artefacts introduced during the digitisation process itself, such as uneven lighting, blurriness, or noise. Although manual restoration by experts is effective, it can be a labour-intensive and costly process that may not be scalable for large archival projects. Consequently, automated restoration approaches have become a necessary alternative.

Related work. Recent advances in computer vision have demonstrated promising capabilities in this domain. Deep learning architectures have been widely adapted for document enhancement and image restoration in general, using Convolutional Neural Networks (CNNs) [3, 4, 5, 6], Generative Adversarial Networks (GANs) [7, 8], transformer-based models [9, 10], diffusion models [11] or recurrent models [12]. Generative models, in particular, are favoured for their ability to produce visually clean outputs from heavily degraded inputs. However, the application of these models to documents with textual data raises a critical safety concern: semantic fidelity. Unlike natural image restoration, where a 'hallucinated' object may be acceptable, document restoration requires absolute precision. A 'hallucinated' character, e.g. a model changing a '3' to an '8' or a 'price' to a 'prize', can significantly alter the meaning of the text, leading to misinformation or misinterpretation of historical records. Despite this risk, the evaluation of document enhancement models remains heavily biased toward perceptual quality. The performance of these models is typically assessed using only standard Image Quality Assessment (IQA) metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [5, 8, 9, 11]. Although these metrics effectively quantify visual fidelity, they do not account for semantic accuracy, meaning that a restored document can achieve a high PSNR score while containing critical textual errors. For example, in [12], the authors have reported that the removal of degradation artefacts is an important step in the restoration process and text extraction, but did not compare the Optical Character Recognition (OCR) results of degraded and restored documents. In another example, the authors of [7] computed the OCR performance on only

*tuch1@uniba.sk
†zuzana.cernekova@fmph.uniba.sk

four instances of degraded-restored document pairs, which could be insufficient to draw meaningful conclusions about the performance of the model in terms of semantic fidelity. More notably, a survey [13] on document enhancement models explicitly excluded OCR metrics from its scope, highlighting a gap in the field regarding the functional safety of these models.

Our contribution. This paper addresses this gap by prioritising semantic safety over visual quality in document restoration evaluation. We investigate the extent to which deep learning models preserve textual integrity by implementing and training three distinct architectures – Pix2Pix [14], DnCNN [4], and DRUNet [3] – on a custom benchmark dataset[1]. We developed a synthetic document generation pipeline to create a dataset that simulates various types of degradation commonly found in documents. To evaluate the models, we introduce a safety assessment framework that categorises OCR errors into safe errors (deletions or data loss) and unsafe errors (substitutions or insertions). Through analysis using both IQA metrics and our proposed safety metric, we provide a comprehensive comparison of the models' performance and safety in document restoration tasks.

2 Synthetic dataset generation

A fundamental requirement for training supervised restoration models is the availability of a large dataset containing pixel-aligned pairs of degraded inputs and clean ground truth targets. While real-world datasets of degraded documents exist, e.g. DIBCO [15] or Labeculae Vivae [16], they typically provide binarized masks rather than clean reconstructions or lack paired ground truth entirely. Additionally, evaluating semantic fidelity requires access to the original text of the documents, which is often unavailable in real-world datasets.

To address the lack of paired ground truth in real-world document datasets, we implemented a custom Python generation pipeline to create a synthetic dataset of degraded images with known ground truth. We selected English and Czech texts from the Project Gutenberg repository[2], ensuring that the models are robust to diacritics. The paper textures and stains were sourced from Texture Labs[3] to simulate realistic degradations.

2.1 Generation pipeline

We implemented a multi-stage pipeline to produce the synthetic dataset. The process is illustrated in Figure 1.

[1]The source code, dataset, training logs and model checkpoints are available at: https://github.com/filippt1/doc_safety/.

[2]https://gutenberg.org

[3]https://texturelabs.org

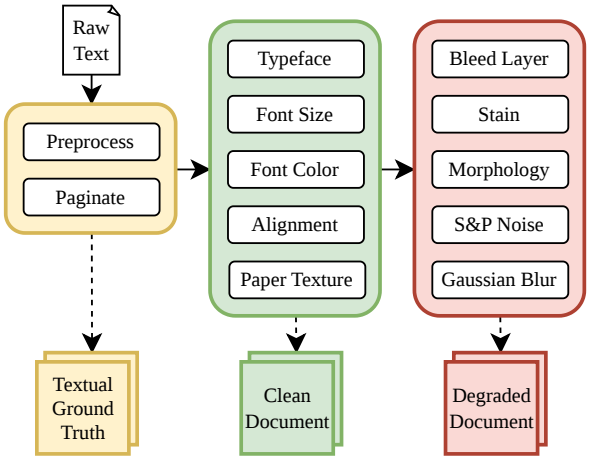


Figure 1: Overview of the synthetic dataset generation pipeline. The process consists of processing the input text, generating a clean document image, and simulating degradations to create the degraded input image.

Generating clean documents. First, the input texts were preprocessed by removing unwanted characters (e.g. '_', '≫') and paginated into segments that fit images of size 256×256 pixels. For each page, we select a random font from a set of standard typefaces (e.g. Times New Roman, Georgia) with sizes ranging from 14pt to 24pt. The text is rendered onto a white background with randomised margins and alignment. To mimic the non-uniformity of real ink, character strokes are not rendered as perfect black, but rather with variations in grey levels. The rendered text is then blended onto a randomly cropped paper texture. This results in clean and legible documents that retain the appearance of paper.

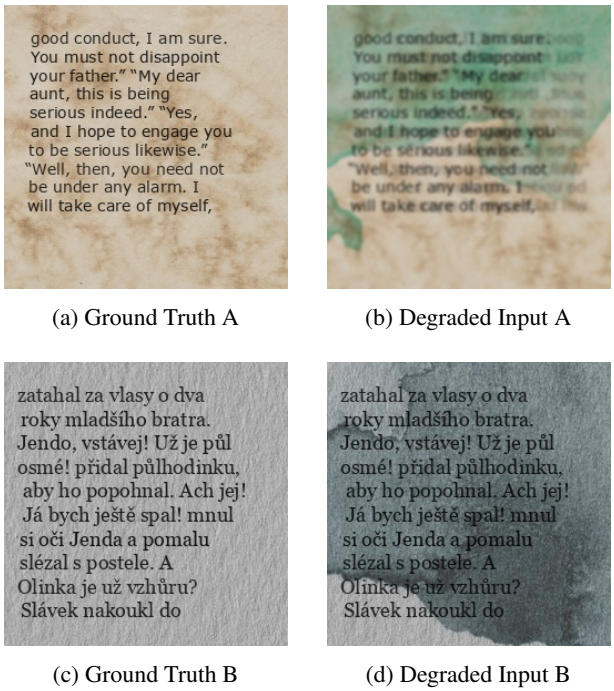
Simulating degradations. The degraded input is generated by applying a stochastic sequence of transformations to the clean document. These transformations include the addition of a 'bleed layer' of text from the following or preceding page (with probability $p = 0.4$), the blending of stains from the sourced dataset ($p = 0.7$) and the introduction of random morphological erosion and dilation ($p = 0.4$). To simulate digitisation artefacts, we inject salt-and-pepper noise ($p = 0.3$) and apply Gaussian blur ($p = 0.7$) with a random kernel size (3×3 or 5×5).

Textual ground truth. Crucially, the pipeline exports the exact text corresponding to the rendered image. This allows for a precise evaluation of semantic fidelity, rather than relying on an OCR engine's interpretation of the clean document.

2.2 Dataset partition

The final dataset consists of generated image pairs (x, y) along with their textual labels. To ensure reproducible evaluation, we partitioned the dataset using a fixed random seed

into training (80%), validation (10%), and testing (10%) sets. In total, we generated ≈ 21000 image pairs, with ≈ 16800 for training, ≈ 2100 for validation and ≈ 2100 for testing. Figure 2 shows samples from the synthetic dataset, illustrating the variety of degradation types simulated.



(a) Ground Truth A (b) Degraded Input A

(c) Ground Truth B (d) Degraded Input B

Figure 2: Visualizing the synthetic dataset. Top row: A sample from the English corpus exhibiting stained paper, text bleeding through, and blur. Bottom row: A sample from the Czech corpus showing stained paper.

3 Methodology

To evaluate the trade-off between perceptual quality and semantic fidelity in document restoration, we utilise three distinct deep learning architectures: DnCNN, DRUNet, and Pix2Pix. These models were selected to represent different approaches in image restoration: discriminative convolutional neural networks and generative adversarial networks. For this study, we reimplemented all three architectures in Tensorflow/Keras and trained them from scratch on our synthetic dataset.

3.1 Discriminative models

DnCNN. We implemented the Denoising Convolutional Neural Network (DnCNN) as proposed by Zhang et al. [4]. Consistent with the original strategy, the network predicts the residual noise map $\mathscr{R}(x)$, obtaining the restored image via $\hat{y} = x - \mathscr{R}(x)$. The architecture consists of 17 convolutional layers, each with 64 filters of size 3×3, separated by Batch Normalisation and ReLU. The model is optimised by minimising the Mean Squared Error (MSE) loss between the degraded-clean image pairs.

DRUNet. To capture more complex degradation patterns, we also implemented the Denoising Residual U-Net (DRUNet) architecture introduced by Zhang et al. [3]. The network integrates Residual Blocks into a four-scale U-Net architecture. Skip connections concatenate features from the encoder to the decoder to preserve spatial details. Each scale processes features through a series of four Residual Blocks, with each block containing two convolutional layers separated by ReLU activations. As with DnCNN, this model is trained to minimise the MSE loss between the predicted and ground truth images.

3.2 Generative model

Pix2Pix. For the generative approach, we implemented the Pix2Pix architecture proposed by Isola et al. [14], a conditional GAN designed for image-to-image translation tasks. The architecture consists of two adversarial networks: a generator and a discriminator. We utilise a standard 8-layer U-Net generator, which features an encoder-decoder structure with skip connections between corresponding layers, resembling the DRUNet architecture without Residual Blocks. The encoder uses Convolution-Batch Normalization-Leaky ReLU blocks, while the decoder uses Transposed Convolution-Batch Normalization-ReLU blocks, with Dropout ($p = 0.5$) applied in its first three layers. The discriminator is a PatchGAN classifier that classifies the input images as real or fake. Its architecture consists of 4 Convolution-Batch Normalization-Leaky ReLU blocks, followed by a final Convolution layer that outputs the real/fake classification. The weights are initialised using random normal initialisation.

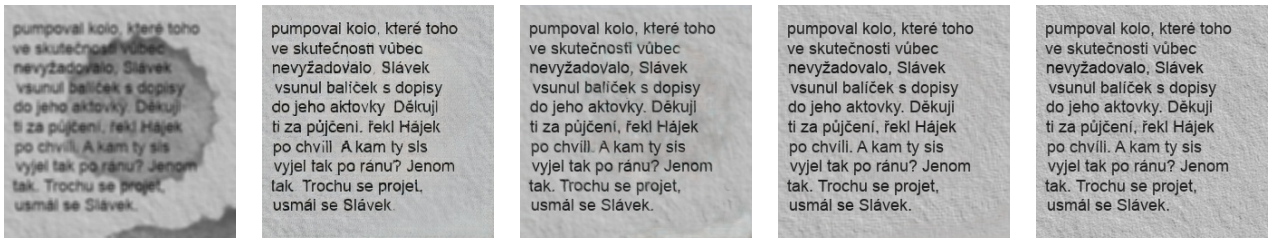
For adversarial training, we use the standard Binary Cross-Entropy (BCE) loss. The discriminator is trained to minimise the classification error for both real and fake images, while the generator is trained to maximize the probability that the discriminator classifies its output as real. To ensure structural fidelity to the ground truth, the generator's objective is combined with a Mean Absolute Error (MAE) loss. We set the MAE weight to $\lambda = 100$, placing a strong emphasis on accuracy at the pixel level.

3.3 Training procedure

The discriminative models were trained for 100 epochs with batch size 16 for DnCNN and 8 for DRUNet, using the Adam optimiser with a learning rate of 10^{-4}. Pix2Pix was trained for 200 epochs with a batch size of 64, using the Adam optimiser for both the generator and the discriminator, with learning rates of $2 \cdot 10^{-4}$. The model checkpoints were saved based on the validation performance: minimum validation loss for DnCNN and DRUNet and minimum validation generator loss for Pix2Pix. The training was performed on an NVIDIA RTX 4080 Super GPU, with

Table 1: Performance comparison of the three restoration models on the test set. DRUNet achieves the best results across all metrics. Note that Pix2Pix shows a large performance drop on EasyOCR.

Model	PSNR	SSIM	MSE	CER (Tesseract)	CER (EasyOCR)
Degraded Input	20.77	0.60	1210.59	0.52	0.47
Pix2Pix	24.02	0.73	423.48	0.14	0.22
DnCNN	26.92	0.82	206.56	0.07	0.15
DRUNet	**32.58**	**0.89**	**58.11**	**0.03**	**0.08**
Clean Target	–	–	–	0.01	0.07



(a) Degraded Input (b) Pix2Pix (c) DnCNN (d) DRUNet (e) Ground Truth

Figure 3: Visual comparison of a degraded input, its enhancements with Pix2Pix, DnCNN and DRUNet, and the target ground truth.

training times of approximately 7 hours for DnCNN, 17 hours for DRUNet, and 4 hours for Pix2Pix.

4 Results

We evaluated the models on the held-out test set of our synthetic dataset. To ensure the reliability of our text extraction metrics, we use two distinct OCR engines, Tesseract and EasyOCR, which have different architectures and training data.

We establish two baselines. The degraded (non-enhanced) input images define the lower bound for the raw legibility and quality of the documents. The clean (ground-truth) target images define the upper bound for the perceptual quality and establish the inherent error rate of the OCR engines themselves.

4.1 Metrics

To provide a comprehensive evaluation, we employ three categories of metrics.

Image Quality Assessment. We use standard metrics to measure the pixel-level and structural fidelity between the reconstructed and ground truth images: Peak-Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) and Mean Squared Error (MSE).

Text Recognition Accuracy. To quantify legibility, we used the Character Error Rate (CER). This metric uses the Levenshtein distance, which measures the minimum number of single-character edits (substitutions, deletions, insertions) required to change the predicted text into the ground truth text. The CER is calculated as $(S+D+I)/N$ where S, D, and I represent the number of substitutions, deletions, and insertions, respectively, and N is the total number of characters in the ground truth. Alignment is resolved by dynamic programming, which prefers a substitution (cost 1) over a deletion followed by an insertion (cost 2). For example, ground truth 'computer' vs. prediction 'cmputors' yields one deletion ('o'), one substitution ('e' to 'o'), and one insertion ('s'). This produces CER $= 3/8 = 0.375$.

Safety Assessment. Because the standard CER aggregates all errors into a single score, we distinguish between errors to assess semantic safety. We categorise deletions as safe errors, as they represent loss of data without introducing false information. In contrast, we categorise substitutions and insertions as unsafe errors (hallucinations), as they introduce false information that appears legible. An unsafe model is one that introduces a significant number of unsafe errors, potentially altering the meaning of the text.

4.2 Image quality and character error analysis

Table 1 summarises the performance in all metrics. DRUNet achieves the best overall results, achieving the highest PSNR and SSIM scores, along with the lowest CER across both OCR engines. On the other hand, the Pix2Pix model significantly underperformed. Although designed to improve the visual quality of images, it scored the lowest

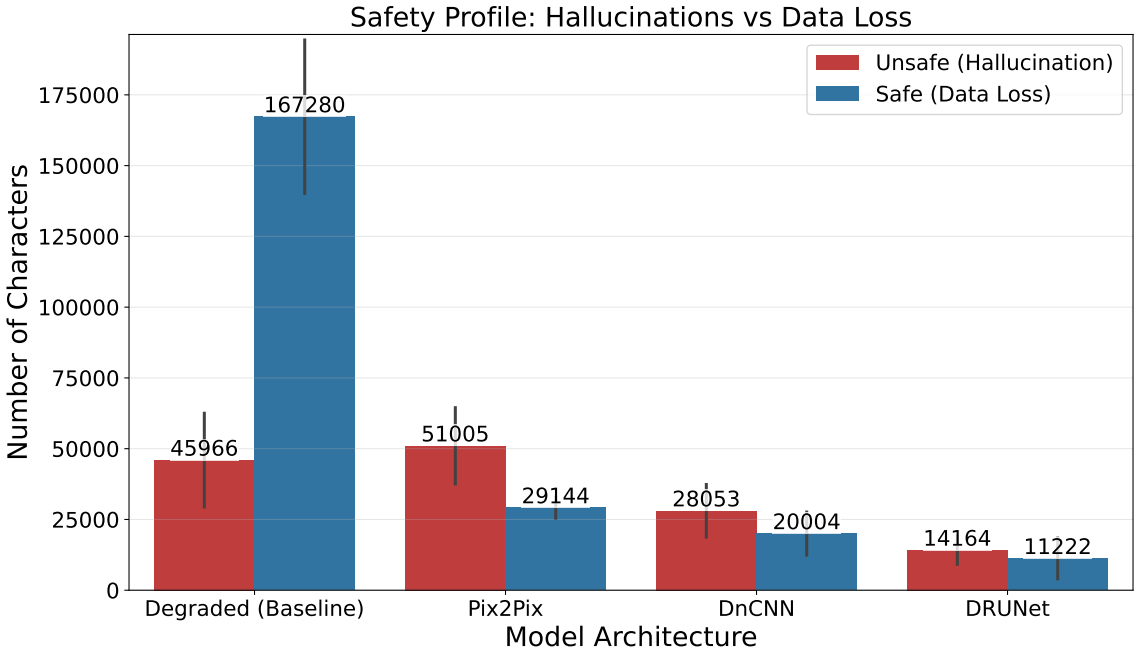


Figure 4: Safety profile analysis showing the breakdown of Character Error Rate (CER) into safe errors (deletions) and unsafe errors (substitutions and insertions) for each model. Pix2Pix introduces higher number of unsafe errors compared to the degraded input, indicating 'hallucination' of characters.

in PSNR and SSIM, and also exhibited the highest CER, particularly with EasyOCR. DnCNN performed well, surpassing Pix2Pix in both image quality and text recognition accuracy, while falling slightly short of DRUNet.

4.3 Visual comparison

In Figure 3, we present a visual comparison of restored documents. The example shows that Pix2Pix tried to restore the image by creating strong edges and textures, but in doing so, it introduced significant artefacts and distortions, which likely contributed to its poor performance in both IQA and CER metrics. DnCNN produced a cleaner image with fewer artefacts, but was more conservative in restoring the text, which resulted in a blurrier output. DRUNet, however, effectively removed the stains and preserved the character details, resulting in a visually clean document that closely resembles the ground truth.

4.4 Safety profile analysis

In Figure 4, we plot the safe and unsafe error rates for each model, along with the degraded input baseline. The degraded input is primarily characterised by safe errors, as degradation renders the text illegible. DnCNN and DRUNet effectively recover these data, while keeping the unsafe errors low. In contrast, Pix2Pix exhibits a dangerous profile, with an increase in unsafe errors compared to the degraded input. In Appendix A, we provide confusion matrices for Pix2Pix and DRUNet evaluated with Tesseract, illustrating the most common character substitutions and highlighting the 'hallucinating' behaviour of Pix2Pix.

In addition, we analysed special characters with diacritics from the Czech corpus, which are particularly prone to alteration by the models due to their visual features. Table 2 presents a detailed breakdown of the models' performance on the 5 most frequent characters with diacritics. We can see that DRUNet consistently outperforms the other models in both accuracy and minimising unsafe errors for these characters, while Pix2Pix shows the lowest accuracy and highest unsafe error rates, often substituting characters with diacritics with visually similar but incorrect characters.

4.5 Discussion

The evaluation results reveal differences in the performance and safety profiles of the three restoration models. These are primarily driven by their architectural differences and training objectives.

Pixel-wise optimization vs. generative modeling. Our analysis shows the risks of generative models in document restoration. While Pix2Pix produces perceptually plausible images, it frequently 'hallucinates' characters, replacing uncertain text with incorrect alternatives. In contrast, discriminative models (DnCNN and DRUNet) trained with pixel-wise loss act conservatively: when faced with uncertainty, they are more likely to leave characters degraded (safe errors) rather than risk introducing incorrect characters (unsafe errors).

Table 2: Performance analysis of DnCNN, DRUNet, and Pix2Pix on the 5 most occurring characters with diacritics. OCR was performed using the Tesseract engine. We can see that the 'í' character is substituted into 'i' in 27.1% of cases by Pix2Pix.

Character	Count	Model	Accuracy	Most Common Error	% Error
í	4501	DRUNet	88.2%	i	9.0%
		DnCNN	72.2%	i	19.0%
		Pix2Pix	57.7%	i	27.1%
á	4075	DRUNet	93.0%	a	5.0%
		DnCNN	84.8%	a	9.0%
		Pix2Pix	76.6%	a	10.3%
ě	3766	DRUNet	94.8%	é	2.5%
		DnCNN	88.0%	é	5.2%
		Pix2Pix	74.3%	é	11.9%
ž	2455	DRUNet	94.0%	Z	1.5%
		DnCNN	88.4%	deletion	4.3%
		Pix2Pix	75.1%	deletion	8.2%
ř	2369	DRUNet	94.3%	deletion	1.6%
		DnCNN	86.8%	deletion	3.5%
		Pix2Pix	73.7%	deletion	7.7%

Global context vs. local details. Comparing the two discriminative models, DRUNet outperforms DnCNN, particularly in removing spatially complex degradations such as stains and bleed-through. As observed in the visual comparison (Figure 3), DnCNN struggled to remove large background stains, effectively treating them as valid document textures. This limitation can be attributed to the shallow receptive field of DnCNN and standard CNNs in general. These architectures process the image in small local patches without capturing broader contextual information. However, DRUNet leverages its U-Net architecture. Its downsampling path enables the model to capture global context and recognize large-scale degradation patterns (such as stains), while the upsampling path ensures the reconstruction of local character details and accurate paper textures. This approach enables DRUNet to achieve the best balance between aggressive removal of degradation and the preservation of the finer textual and textural structures.

5 Conclusion

The automated restoration of historical documents presents a unique challenge in which the enhancement of visual quality needs to be balanced with the accurate preservation of the original text. In this paper, we evaluated the functional safety of three deep learning models – DnCNN, DRUNet and Pix2Pix – on a synthetic dataset of degraded text documents. By introducing a safety assessment framework that distinguishes between safe errors (deletions) and unsafe errors (hallucinations), we showed risks that are often overlooked by traditional image quality metrics.

We show that while generative models like Pix2Pix can greatly improve the visual quality of degraded documents, they could potentially be unsafe for archival purposes. The tendency of the generative model to 'hallucinate' characters and remove diacritics poses a significant risk of altering historical records. On the contrary, discriminative models trained with pixel-wise objectives proved to be more effective and conservative while enhancing the text. Among them, DRUNet proved to be the superior architecture, leveraging its U-Net structure to efficiently remove document degradations while preserving the original document content.

We conclude that, for digitisation of historical archives, the field should shift from assessing the performance of restoration models purely based on image quality metrics, to assessing also the integrity of the models and their risk of altering the content of the documents.

Future work could include evaluating a wider range of architectures, including transformer-based models and diffusion models, which have shown promise in image restoration tasks. Additionally, developing a model that explicitly incorporates OCR feedback into the restoration process could further enhance the semantic fidelity of restored documents. Applying the proposed safety assessment framework to real-world historical document datasets would provide valuable insights into the practical implications of these models in archival digitisation projects.

Acknowledgements

This work was supported by the APVV-23-0250 project.

References

[1] Sílvia Sequeira and Flavia Pinzari. *Degradation, Remediation and Protection of Library Materials*, pages

15–40. Archetype, 06 2022.

[2] Islam El Jaddaoui, Hassan Ghazal, and Joan Bennett. Mold in Paradise: A Review of Fungi Found in Libraries. *Journal of Fungi*, 9:1061, 10 2023.

[3] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-Play Image Restoration with Deep Denoiser Prior, 2021.

[4] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, July 2017.

[5] Sarra Hamza, Rafik Menassel, and Chawki Djeddi. RSEUNet: residual squeeze and excitation U-Net for restoration and binarization of historical documents. *Multimedia Tools and Applications*, 84:48511–48534, 08 2025.

[6] Darayut Nhem and Chhim Bunchhun. Khmer Historical Document Image Restoration Using U-Net's Variants. *Insight Cambodia Journal of Basic and Applied Research*, 7, 12 2025.

[7] Mohamed Ali Souibgui and Yousri Kessentini. DE-GAN: A Conditional Generative Adversarial Network for Document Enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1180–1191, March 2022.

[8] Raphaela Heil and Fredrik Wahlberg. Restoration of Archival Images Using Neural Networks. *Digital Humanities in the Nordic and Baltic Countries Publications*, 4:79–93, 10 2022.

[9] Mingxian Li, Hao Sun, Yingtie Lei, Xiaofeng Zhang, Yihang Dong, Yilin Zhou, Zimeng Li, and Xuhang Chen. High-Fidelity Document Stain Removal via A Large-Scale Real-World Dataset and A Memory-Augmented Transformer, 2024.

[10] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image Restoration Using Swin Transformer, 2021.

[11] Fangmin Zhao, Weichao Zeng, Zhenhang Li, Dongbao Yang, Binbin Li, Xiaojun Bi, and Yu Zhou. UniDocDiff: A Unified Document Restoration Model Based on Diffusion. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 8204–8213, New York, NY, USA, 2025. Association for Computing Machinery.

[12] Yashowardhan Shinde, Kishore Kulkarni, and Sachin Kuberkar. EraseNet: A Recurrent Residual Network for Supervised Document Cleaning, 2023.

[13] Lucas N. Kirsten, Ricardo Piccoli, and Ricardo Ribani. Evaluating Deep Neural Networks for Image Document Enhancement, 2021.

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-Image Translation with Conditional Adversarial Networks, 2018.

[15] Ioannis Pratikakis, Konstantinos Zagoris, Xenofon Karagiannis, Lazaros Tsochatzidis, Tanmoy Mondal, and Isabelle Marthot-Santaniello. ICDAR 2019 Competition on Document Image Binarization (DIBCO 2019). In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1547–1556, 2019.

[16] Alberto Campagnolo, Erin Connelly, and Heather Wacha. Labeculæ Vivæ: Building a Reference Library of Stains for Medieval and Early Modern Manuscripts. *Manuscript Studies A Journal of the Schoenberg Institute for Manuscript Studies*, 4:401–416, 11 2019.

A Character confusion matrix

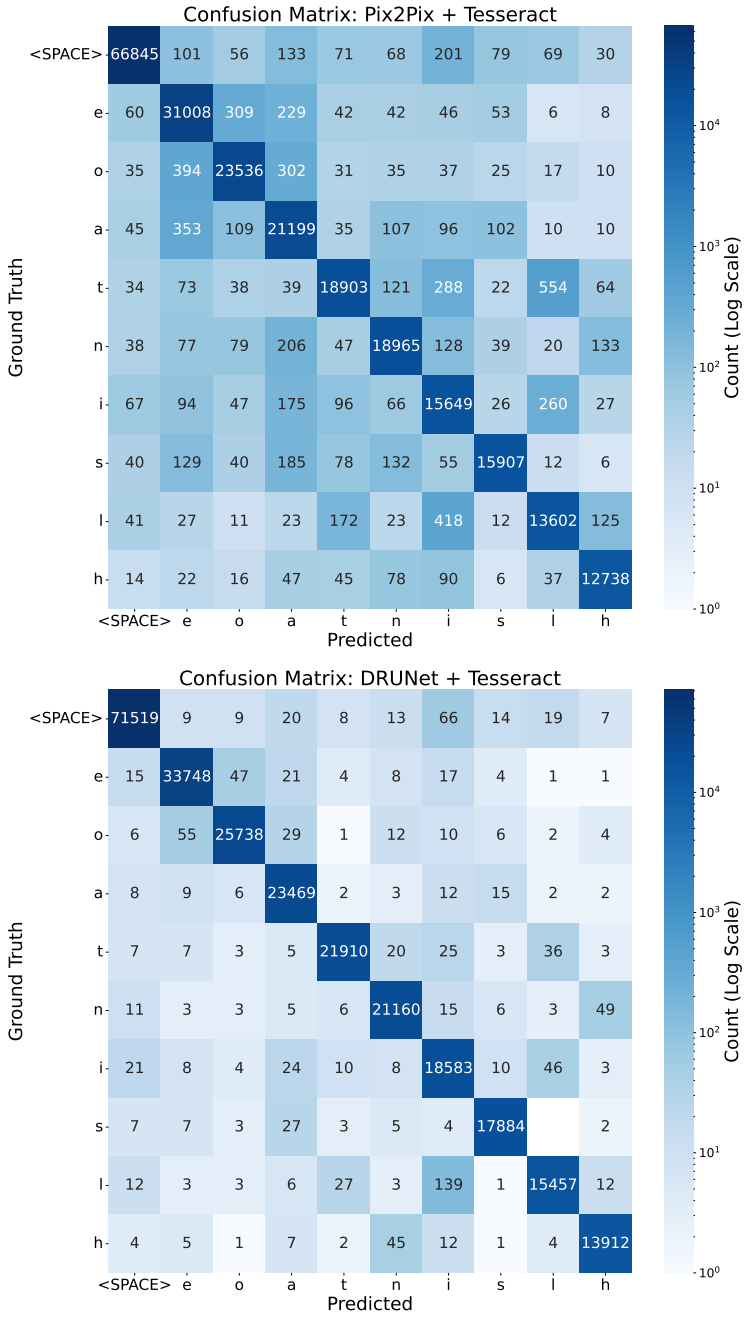


Figure 5: Comparison of character confusion matrices for Pix2Pix and DRUNet evaluated with Tesseract. Note the 'hallucinating' behaviour of Pix2Pix, which frequently substitutes characters with visually similar but incorrect alternatives.