Image-to-Image Translation Approach for Page Layout Analysis and Artificial Generation of Historical Manuscripts

Chahan Vidal-Gorène, Jean-Baptiste Camps

Image-to-Image Translation approach for Page Layout Analysis and Artificial Generation of Historical Manuscripts

Chahan Vidal-Gorène[1 [0000−0003−1567−6508]] and Jean-Baptiste Camps[1 [0000−0003−0385−7037]]

École nationale des chartes, Université Paris, Sciences & Lettres, 65 rue de Richelieu, 75002 Paris {chahan.vidal-gorene,jean-baptiste.camps}@chartes.psl.eu

Abstract. Document layout analysis is essential in Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR), especially for historical and low-resource scripts. This study explores a novel data augmentation technique using Generative Adversarial Networks (GANs) to generate realistic document layouts from semantic masks, enhancing layout analysis without increasing human annotation effort.

Our lightweight pipeline, tested on historical manuscripts (Latin, Arabic, Armenian, Hebrew), newspapers, and complex document layouts, shows that GAN-generated layouts are convincing and difficult to distinguish from real ones, even for paleographers. This method significantly boosts data augmentation, yielding a 3 percentage point improvement in layout analysis metrics (precision, recall, mAP), and a 12 point increase in precision and recall for damaged documents. Additionally, masks with character information enhance image quality, boosting text recognition performance.

Keywords: GAN · Layout Analysis · Semantic Classification · Data Augmentation · Handwritten Text Recognition

1 Introduction

Layout analysis is a core task in the analysis and understanding of documents, that involves the detection and annotation of physical areas on the source material; its numerous applications range from document categorisation to text recognition, and is generally intended as a pre-processing step [4], that will affect the results and accuracy of all ulterior treatments. Many approaches exist, the most common being to separate the detection of layout areas (e.g., text columns, marginal notes, illustrations, figures, etc.) and of text lines: the page is first divided into regions, then the lines are detected within the relevant regions.

These steps are largely covered by state-of-the-art systems [9,10], in particular to detect baselines and propose a semantic classification of text regions (e.g. marginal note, title, etc.). Many datasets offer a two-level annotation of their

contents [25], and the use of an ontology like SegmOnto [12] allows their pooling in order to enlarge the training datasets and lead to more versatile models.

In the case of historical documents, the task to overcome is more complex due to the very wide variety of layouts, physical support and page preparation, size of scripts, handwritten traditions, irregularities, damage to documents (e.g. worm holes, burnt manuscripts,...) or even scan quality. The creation of datasets for complex documents comes up regularly, betting for example on the variety [9], or on a family of manuscripts [19]. Despite the availability of increasingly versatile layout analysis models that enable document pre-annotation within mainstream platforms [18,20,37], any historical document processing project must even today consider a massive annotation of layouts in order to overcome a given corpus with a specialised model. Document annotation, even semi-automated, remains a time-consuming task, and incompatibilities between datasets (see section 4) delay the arrival of very versatile models for historical documents. We therefore propose a new method of data augmentation, for the generation of artificial pages with complex layouts emulating historical documents, and evaluate it on several datasets and use cases. In particular, we envision the case of historical manuscripts, such as those damaged by fire.

2 Related work

Recent efforts in layout analysis are integrating the detection of text regions and text lines into a singular task, primarily through the use of transformers [5,24,41], which achieve scores comparable to traditional state-of-the-art methods. However, these transformer-based approaches often require a large volume of data. Alternative methods include simple CRNN layers [19], effective in clear cases but struggling with closely situated regions of the same type. In contrast, U-net models [14,22,32,37] have led recent competitions [10] with less data dependency. Object detection strategies [8] also show promise, frequently surpassing other methods by leveraging extensively pre-trained models [24].

Generative adversarial networks (GANs) are increasingly employed for qualitative dataset augmentation [33] and have begun to influence document analysis with the integration of transformers in scientific document processing [1,30]. These are trained on extensive datasets such as scientific articles [28,42] and have been used to generate both printed [6] and handwritten layouts [22] on varied corpora, in particular modern handwriting [31]. Diffusion models have also been explored for generating scientific documents without text constraints, but only produce low-resolution results [35]. Despite the challenges posed by historical documents due to high noise levels, transformers are becoming feasible for these as data scarcity issues are addressed, improving HTR line performance [3]. GANs, with or without transformers, have been primarily evaluated for reducing the Character Error Rate (CER) of HTR models by generating fake text lines [11,36,40], achieving notable success in both Latin and non-Latin scripts [33]. However, their application for layout analysis in historical documents has been very limited.

While no definitive metric exists to assess the quality of GANs for text image generation (see subsection 5.2), these technologies are capable of producing qualitatively convincing layouts and text lines, except for figures and graphs. Most methods rely on a constrained approach where GANs generate layouts from semantic region coordinates provided in an XML file. This results in well-structured outputs, such as printed scientific articles or handwritten tables. A style transfer approach has also been tested on historical documents [39], applying a handwritten style to a printed template. While the clarity of template text boxes leads to an effective imitation, it fails to represent the deterioration found in historical documents. The same applies to complex documents like newspapers.

3 Proposed method

We propose a method of artificial page generation, that accounts both for the constraint of emulating existing and relatively specific historical layouts, and being able of generating entirely new artificial pages, without over-specifying their composition. For this, we propose a method that relies on constrained image-to-image translation, generating artificial pages from input layout masks of regions and baselines.

The objective of this approach is to propose a framework constrained to GANs, utilizing explicit minimalist semantic information (regions, lines, or characters). A historical handwritten document style mapping is then applied to these semantic areas, allowing the GAN to focus solely on reproducing the writing aesthetic and the background. Our approach reuses the implementation provided by Pix2Pix [17]. The GAN establishes a relationship between image pixels and generated masks for object layouts. Generating false layout masks and transforming them into fake pages with known coordinates, inspired by the creation of maps or building facades [17], is less time-consuming than annotating documents. For historical documents, Pix2Pix has already been used to synthesise data for palimpsests reconstruction [21].

For each dataset, given sufficient data (see section 4), we generate semantic masks for text regions and baselines, as illustrated in Figure 1. This process allows for a dataset of masks equivalent to the image dataset. We are currently exploring both text region semantic masks and line masks. We also apply the same method at the character level to evaluate its transfer to a more qualitative task.

At the image generation stage, a semantic mask generator creates random masks, saving object coordinates in an XML file. Our GAN, pre-trained on a document type, then generates a synthetic image. This results in an XML of semantically tagged coordinates and an image (see Figure 1). In details, we give to Pix2Pix random patches of 512x512px, cropped from a rescale image of 2048px height, with the Unet-256 model Pix2Pix and 300 hidden-size for the Generator. The style is unconstrained; only the semantic spatial information's position is dictated by the generator.

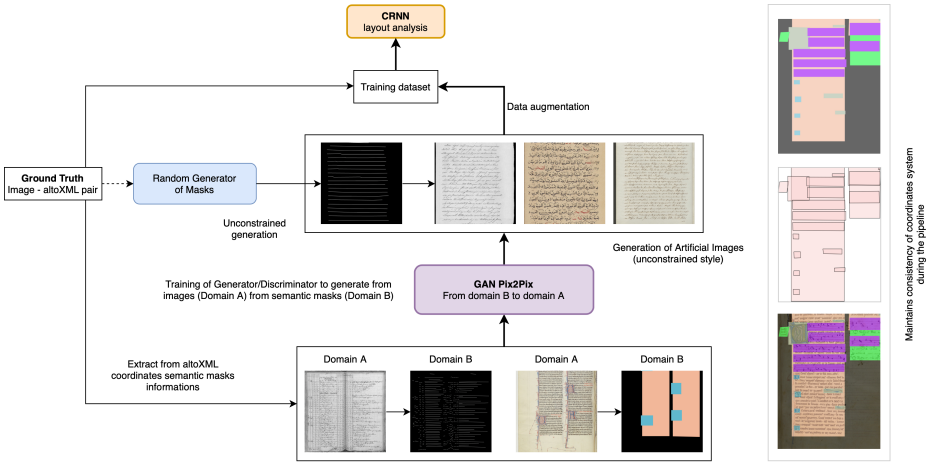


Fig. 1. Training pipeline using Pix2Pix at Region level and Baseline level

After data creation, we train a CRNN for layout analysis to verify the potential of this method to create gains in the accuracy of layout analysis tasks.

For generating images with explicit textual information, data augmentation occurs dynamically throughout the recognizer's training. This approach involves pre-training the recognizer for several epochs, then predicting on its training data to achieve a Character Error Rate (CER) under 1%. This prediction helps obtain character coordinates for generating varied random mask compositions.

The resulting images, constituting up to 25% of the training data, provide additional data augmentation without stylistic constraints, though mask dimensions may suggest a style to the GANs (see Figure 2).

Image-to-image translation has demonstrated robust performance across many real-world applications. Despite its age compared to newer GAN models, supervised methods generally outperform unsupervised ones in image-to-image translation, even though advanced architectures like StyleGAN combined with contrastive learning techniques also show promising results [16,27]. These approaches, however, are not evaluated in our experiments.

4 Datasets

We conducted our main experiments on layout on three types of data:

1. **Manuscripts in Latin scripts**: CREMMA Medieval [29] with manuscripts in Old French and Latin, as well as cBAD 2017 Simple & Complex Track [9], that contains samples from documents written between 1470 and 1930, and coming from nine different European archives in Belgium, England, Finland, Germany, Italy and Switzerland [13].
2. **Manuscripts in non-Latin scripts**: RASAM for Arabic Maghribi [38], BADAM for Arabic scripts [19], and BiblIA for medieval Hebrew [34].

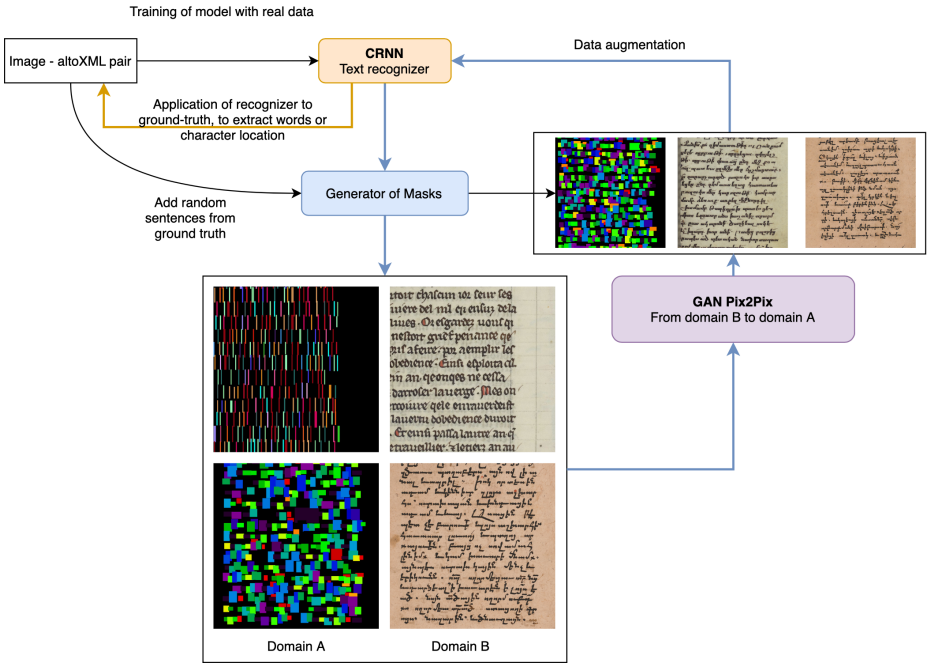


Fig. 2. Training pipeline using Pix2Pix at the character level

3. **Complex printed documents in Latin types** from the NewsEye READ dataset of contemporary French newspapers [23].

Experiments were conducted separately on each dataset to maintain control over the style generated and accommodate the varying levels of annotation detail provided for baselines and text regions.

The CREMMA Medieval and RASAM datasets served as our primary evaluation sources, with the CREMMA Medieval dataset offering comprehensive annotations using the SegmOnto ontology. Conversely, the BADAM, BiblIA, and NewsEye datasets were employed to test the viability of our approach on non-Latin scripts and printed documents with complex layouts due to their partial annotations (incomplete regions and lines).

The cBAD dataset, while included, presented challenges due to imprecise and insufficiently detailed annotations, rendering it less effective for our specific tasks. A detailed summary of each dataset's characteristics is provided in Table 1.

Complementary experiments on character and layout analysis tasks were conducted using the CREMMA Medieval, textscbadam, textscrasam, textsccbad Simple Track, and NewsEye datasets, chosen for their representation of diverse tasks. Additionally, a specialized Armenian dataset, not available in open access, was utilized to quantitatively evaluate the results and highlight specific paleographic patterns produced by Pix2Pix.

Table 1. Composition of datasets

Dataset	Baselines	Regions	Annotation	Ontology	Pages
Manuscripts					
Latin scripts					
CREMMA Medieval	✓	✓	full	SegmOnto	600
cBAD 2017 Simple	✓	✓	full	cBAD	703
cBAD 2017 Complex	✓	✓	full	cBAD	1060
Non-Latin scripts					
RASAM (Arabic Maghribi)	✓	✓	full	custom	300
BADAM (Arabic, mixed)	✓	-	partial	none	400
BiblIA (Hebrew)	✓	✓	partial	none	132
Printed documents					
Latin types					
NewsEye	-	✓	full	custom	630

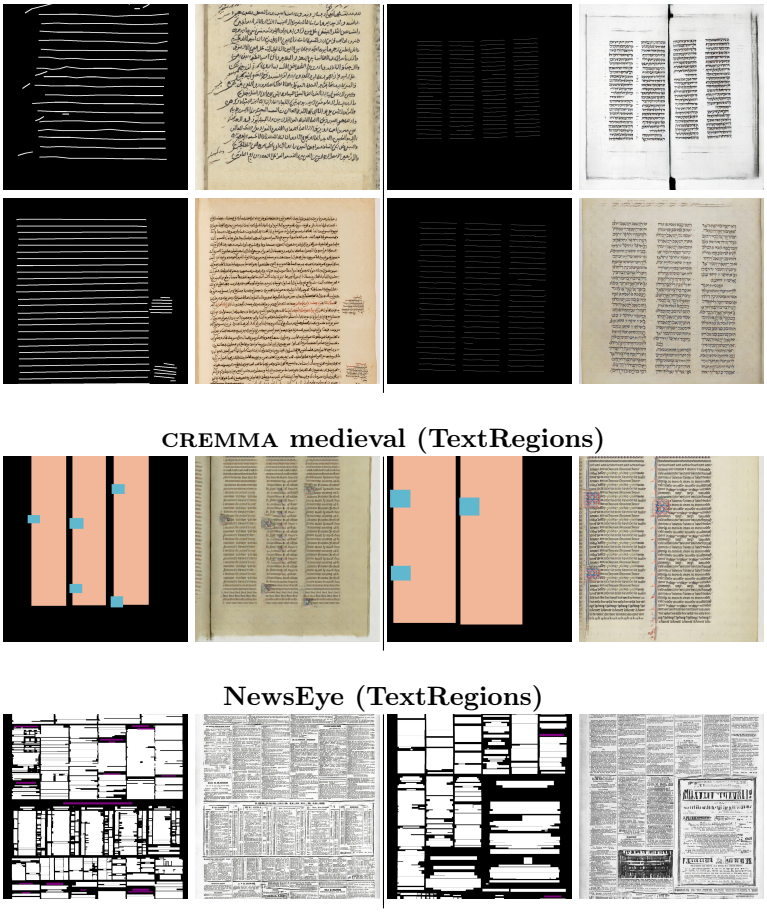
5 Results

Example of generations of artificial pages output by our model are presented in Table 2 for Lines and TextRegions generation, as well as in Figure 3 and Figure 4 for Lines and Characters generation. In our experiments, neither the diversity of objects in the dataset, nor the number of training samples appeared to significantly impact the model's ability to generate convincing images (e.g. while the CREMMA-Medieval dataset contains several pages of each source manuscript, the cBAD complex is composed only of isolated pages from quite diverse documents, without any obvious effect on the model generations).

5.1 Qualitative assessment

The results across various datasets display a significant level of verisimilitude, particularly when evaluated holistically without focusing intensely on the textual content or the minutiae of the generated images. At the layout level, the models reproduce structural information that closely mirrors actual documents.

Table 2: Output artificial images, generated by our model, based on the training on various datasets. The input mask is presented on the left, and the resulting generation on the right. The examples displayed here are sometimes strongly realistic on the basis of a general expert assessment



In the CREMMA dataset, for instance, the generated layouts accurately reflect the typical two or three-column configurations found in medieval French and Latin manuscripts. Additionally, the models are able to capture and emulate to a point finer details such as secondary decorations including initial letters, pen-flourishing extending into margins, and detached verse initials situated in their own columns. Major decorative elements like illuminations and painted miniatures are also replicated (Table 2). The pages generated solely from baseline data interestingly allow the model to infer semantic units such as columns and adequately fill spaces designated for decorations, without explicit semantic information in the mask regarding the presence of such a type of zone. These baseline-derived layouts present a more irregular, authentic appearance, which is characteristic of such manuscripts, as opposed to those generated from semantic regions that tend to appear overly regular due to the geometric shapes of the masks used. Furthermore, the model demonstrates the ability to emulate different script styles and hands, offering variations of the Gothic *Textualis* script with varying degrees of roundness and formality. However, the models still

struggle to produce convincing illuminations or coherent text, often repeating the same words multiple times within a line, due to the lack of an integrated language model. This issue is particularly noticeable when more empty pages with minimal text are provided.

Similar patterns are observed in the cBAD dataset (Table 2), where the model successfully interprets elements like table layouts. The scripts' cursiveness is well-rendered, although the content often appears gibberish-like upon closer inspection, contrasting with the more consistent CREMMA results. This may be due to the cBAD dataset's greater diversity in script types and its broader historical scope.

The non-Latin datasets (Table 2) show the model's capability in rendering complex manuscript layouts with varying text orientations and marginal notes, as well as the general decor and style of manuscripts, such as Arabic ones. The generated texts, while mimicking authentic letter forms and patterns, remain nonsensical. Notably, in the BiblIA dataset, even without annotations for marginal notes, the model attempts to fill these areas, although not as textual content.

The results from newspaper layouts (Table 2) also show a commendable level of credibility, reproducing complex elements such as columns, titles, and article separators effectively. However, due to the small size of input images, the text often does not materialize as coherent letters—a limitation potentially addressable with higher resolution images.

In summary, while the generated images from regions and baseline masks lack textual content (or, at least, textual content that is semantically valid, and goes beyond filler text), their structural fidelity to medieval handwriting styles is impressive, albeit distinguishable from genuine manuscripts upon human evaluation. This study also explores the impact of the absence of textual content on layout analysis effectiveness. At this step, we do not keep the approach using Text Region masks, less accurate than the one using Baseline masks.

For images derived from characters masks, the outcomes are highly convincing, particularly for Armenian scripts, where distinguishing between authentic and synthetic images is challenging, except for the initials, especially decorated ones, and intonational signs—which are not recognized by the HTR and thus absent from the masks—and decorative elements. But results imitate perfectly Armenian manuscripts, no matter the script considered such as *erkat'agir* (Capital script), *bolorgir* (bicameral non-cursive script) and *notrgir* (cusrive script). In Latin scripts, the model struggles with accurately generating the diverse and often underrepresented abbreviative signs in the datasets, but the result is readable and very accurate.

5.2 About the FID metric

To evaluate the quality of images generated by Pix2Pix, we use the Frechet Inception Distance (FID) [15], calculated from features extracted by an Inception V3 model trained on ImageNet. This metric assesses the noise level in generated images compared to real reference images. Notably, when real image sets are

compared among themselves using FID, scores such as 18.75 for Mix BL, 10.43 for HYE char, and 19.33 for LAT char are observed, suggesting these as new benchmarks for generated images due to the limitations inherent in the Inception model's training.

Throughout the training of three models (Figure 3) the FID scores remained high, even post-training. Incorporating the aforementioned real image FID scores as reference values, LAT Char and Mix BL align closely with these targets, while HYE Char achieves a lower FID of 10. This discrepancy may stem from challenges in generating illuminations, where the absence of a proper mask sometimes leads to mere attempts at drawing, despite the generated text's high quality.

Qualitatively, by iteration 1000, the text for HYE Char and LAT Char becomes highly legible, and by iteration 1500, it becomes difficult for a non-expert to distinguish these from real documents. Mix BL, achieving a cleaner appearance by iterations 1500 and 2000, remains identifiable as GAN-generated due to minimal constraints in the generation process and the lack of detailed textual information. However, this does not significantly impact the tasks related to layout analysis (see subsection 5.3).

These findings highlight FID's limitations when applied to GAN-generated handwritten documents. While effective for evaluating global noise and suitable for manuscripts that focus on line generation, FID does not capture finer textual details. Therefore, enhancing FID assessments with readability criteria—either through a Character Error Rate (CER) threshold, if ground truth and a robust HTR model are available, or by defining thresholds for quality assessments [7]—is crucial to accurately evaluate text generation nuances. Parallel training of a recognizer is generally carried out [11,36,39]. This limitation necessitates the development of a new metric that integrates both visual quality and textual readability, providing a more comprehensive assessment of the generated outputs. To this end, we introduce a hybrid quality score (Q), defined as:

$$Q = \alpha \cdot (1 - \text{Norm}(F)) + \beta \cdot \text{Norm}(R) \tag{1}$$

with:

$$\text{Norm}(F) = \frac{F - F_{min}}{F_{max} - F_{min}} \in [0, 1], \text{ and } \text{Norm}(R) = \frac{R - R_{min}}{R_{max} - R_{min}} \in [0, 1] \tag{2}$$

where F is the FID score, R represents the readability score based on the thresholds good (CER $\in [0, 10]$), acceptable (CER $\in [10, 25]$), bad (CER $\in [25, 50]$) and very bad (CER $\in [50, 100]$) [7] converted into a numerical scale, $\text{Norm}(F)$ and $\text{Norm}(R)$ are the normalized values of FID and readability, respectively, ranging from 0 to 1. The parameters α and β are weights that signify the relative importance of each component—visual quality and readability.

In Equation 1, $(1 - \text{Norm}(F))$ inversely scales the FID score to align it with the direct proportionality of higher scores indicating better performance, which is consistent with the readability score. This approach allows the comprehensive evaluation of GAN outputs by quantifying both the aesthetics of the manuscripts and the clarity and legibility of their textual content, addressing a crucial gap

in existing evaluation methodologies for text-containing images generated by GANs.

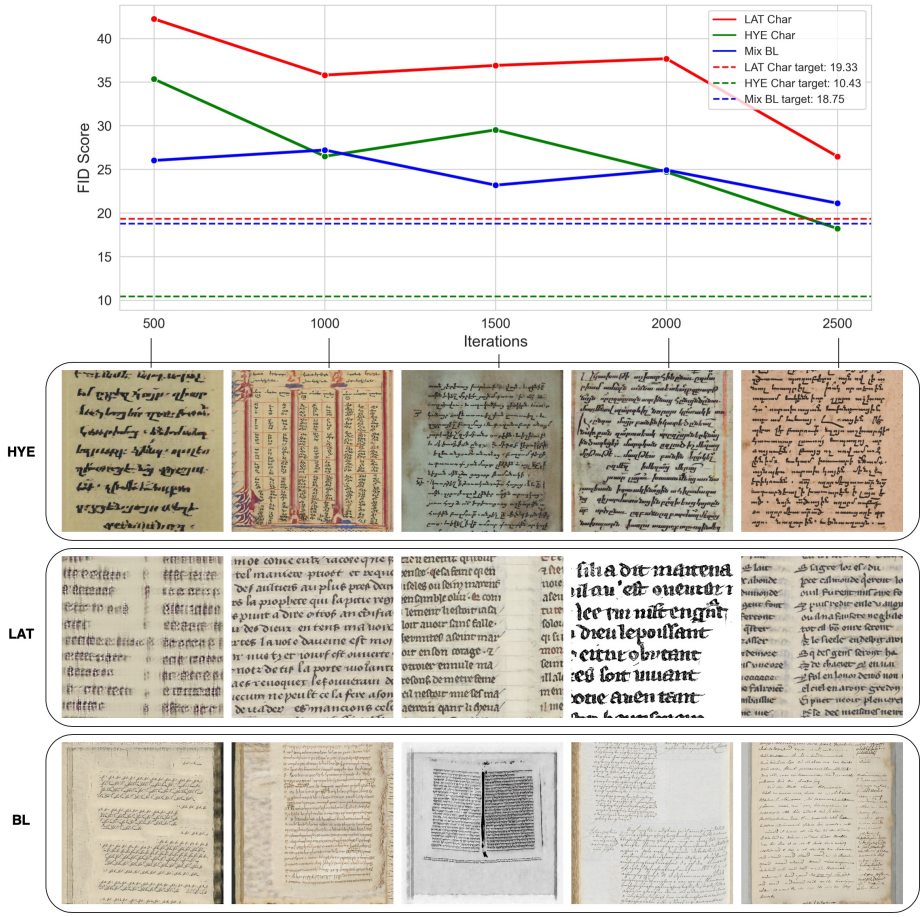


Fig. 3. Evolution of the FID value and corresponding artificial images, generated by the model, for each iteration, for three different datasets, Armenian characters (HYE char), Latin characters (LAT char), and a mixed baseline dataset (Mix BL, including data from CBAD, CREMMA and BADAM)

Applied to HYE Char, the hybrid quality score yields the results reported in Table 3.

We observe significant progress between iterations 500 and 1000, and further improvement from iterations 1000 to 1500, even though the FID scores for iterations 500 and 1000 are comparable. Notably, iteration 1500 is penalized by a higher FID of 29.52 (as opposed to 26.49 and 26.01) due to the generation of a stained background—likely the reproduction of a water stain in the manuscript.

Table 3. Hybrid Quality Score Outcomes for HYE Char. Here, $F_{min} = 0$ represents the lowest value of FID indicating a real image, $F_{max} = 200$ is the threshold for non-credible GAN outputs, $R_{max} = 100$, and $R_{min} = 25$. Both α and β are set to 0.5, giving equal importance to visual quality and readability.

Iteration	FID	Readability	Q Score
500	26.01	Poor	0.435
1000	26.49	Acceptable	0.600
1500	29.52	Good	0.926
2000	24.69	Good	0.938
2500	18.21	Excellent	0.955

This effect, while entirely credible and relevant, is treated as noise in FID calculations, despite the presence of damaged images in the real image set. The iterations 1500 and 2000 exhibit very different FIDs but equivalent readability, which minimally impacts the Hybrid score. Although the definition and parameter weighting of this score may require refinement, the metric as defined here underscores the relevance a a specific metric tailored to the evaluation of GAN-generated written or handwritten documents.

5.3 Benefits on Layout Analysis

To validate our method, we compared baseline detection results on classic datasets, specifically BADAM and cBAD Simple Track [9], using a CRNN initially developed for BADAM [19]. Data augmentation was applied exclusively in the training set, where we dynamically generated 500 new images per training epoch, replacing the previous batch with in-domain constructed masks based on the dataset.

The results are summarized in Table 4, showing the Precision (P) and Recall (R) percentages for various models of the cBAD competitions.

These results illustrate a significant improvement in performance across both the diverse dataset of cBAD Simple Track and the more complex BADAM dataset. While the augmented CRNN model does not yet surpass the top state-of-the-art models on the cBAD Simple Track, the training pipeline of a CRNN combined with Pix2Pix is notably simpler and computationally less demanding, offering greater flexibility. A key takeaway from this experiment is that despite the lack of textual information (random shapes that imitate manuscript) and the presence of margin noise (blurring effect) in the baseline GAN-generated images, these do not impede the CRNN's performance when applied to real images. A CRNN model trained solely on fake images naturally faces more challenges, yet the recall results are commendable, achieving 82.8% on cBAD and 88.3% on BADAM—outperforming the 2021 U-net version [37] and nearly matching the standard CRNN.

The heterogeneity of annotations in classic datasets—such as differences in annotation strategies by paragraph versus column of text, and variability in the completeness of annotations—complicates direct comparisons at the text

Table 4. Baseline Detection Results on cBAD Simple Track (ICDAR 2017) and BADAM

Model	Precision (%)	Recall (%)
cBAD Simple Track (ICDAR 2017)		
dhSegment [26]	94.3	93.9
ARU-Net [14]	**97.7**	98.0
Vision U-net	95.1	95.3
CRNN real	94.4	96.6
CRNN fake	69.3	82.8
CRNN + augment 500	97.4	**98.6**
BADAM		
Vision U-net	91.32	85.75
CRNN real	94.1	90.1
CRNN fake	68.1	88.3
CRNN + augment 500	**96.2**	**91.9**

region level. For instance, some datasets like BiblIA may not include marginal regions in annotations (see Table 1). We conducted experiments on medieval CREMMA, RASAM, and NewsEye (newspapers), employing both CRNN and YOLO v8 models, with and without generated data, applying a dynamic data augmentation strategy.

Despite the application of advanced models and augmentation techniques, the results reveal limited, if any, gains in region detection for both CRNN and YOLO. Notably, there is a significant performance drop in complex datasets like NewsEye, characterized by a high density of regions. The GAN-generated regions often exhibit blurriness and imprecision, and the tendency for the network to generate non-textual information at the exact edge of the mask gives a very artificial appearance to the output (see subsection 5.1). Additionally, the GAN fails to assist the CRNN in distinguishing closely situated regions of the same type, often leading to their amalgamation into a single detected region.

One of the interesting use cases of this approach lies less in the augmentation of data for clean and already well-covered documents than in the creation of qualitative data for a very under-resourced target document (due to the script, the support, etc). A prime example is the manuscript *Torino, Biblioteca Nazionale Universitaria*, L.II.14, which was severely burned in the 1904 library fire. The challenge of extracting content from such a degraded document highlights difficulties in layout analysis. To address this, we experimented with generating damaged layouts to enrich the training dataset, which was qualitatively evaluated at the beginning of this paper. Figure 4 presents an original image from the manuscript, and an artificial generation imitating this manuscript using a semantic mask of baselines.

Figure 5 shows that models trained with this augmented dataset, where generated data replaced up to 25% of the real images per epoch, exhibited a significant improvement in recall for detection tasks. Notably, there was a marked

Table 5. Region Detection Results on CREMMA Medieval, RASAM, and NewsEye

Model	Precision (%)	Recall (%)
CREMMA Medieval		
CRNN real	91.4	89.6
CRNN fake	32.3	37.9
CRNN + augment 500	91.8	94.6
YOLO v8 real	97.2	96.6
YOLO v8 fake	69.3	74.8
YOLO v8 + augment 500	**97.4**	**97.7**
RASAM		
CRNN real	93.4	92.6
CRNN fake	43.2	21.8
CRNN + augment 500	93.8	93.2
YOLO v8 real	**98.1**	97.9
YOLO v8 fake	71.1	77.4
YOLO v8 + augment 500	98.0	**98.3**
NewsEye		
CRNN real	67.4	58.2
CRNN fake	21.1	19.4
CRNN + augment 500	56.2	54.2
YOLO v8 real	**91.7**	**78.2**
YOLO v8 fake	75.8	73.4
YOLO v8 + augment 500	81.4	76.8

decrease in the number of lines fragmented into small pieces compared to models trained solely on real data.

6 Conclusion and further research

Our current results show the ability of our approach to generate artificial pages with layouts that emulate historical documents such as ancient manuscripts in a credible way for an expert, and to significantly improve layout analysis tasks for documents, especially underrepresented documents with rare layouts, noisy or otherwise damaged. Surprisingly, the approach based on input masks of baselines outperforms approaches based on semantic regions masks.

It is surprising that this architecture, focused on lines and layouts, is actually able, in the absence of a language model or an input text, to generate on occasions realistic looking text, despite its nonsensical nature. Another good surprise is the ability of a fully line-based approach, without the encoding of any semantic zones, to still emulate different types of areas in the layouts (e.g., text, decoration, illumination, titles, etc.).

It could be argued, in that case, than the line-based approach is sufficient, and that the area-based approach is unnecessary, especially given the fact that many datasets do not include relevant information. On the other hand, it is easier and faster for a human to create a handful of fake masks of areas, for on-demand page generations, than a complete set of lines.

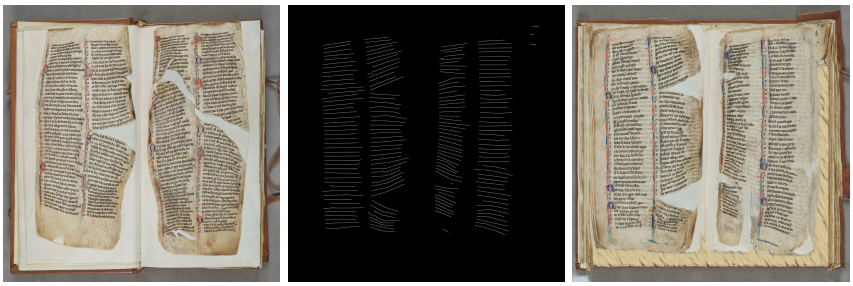


Fig. 4. Generation of damaged document. From Left to Right: Original document (real image), a generated mask, and an artificial image

Fig. 5. Results in layout analysis on a burnt manuscript (Torino, BNU, L.II.14), using only real data in training (on the left) and with data-augmentation using GAN (on the right). Red and green lines indicate indifferently the baselines recognised by the layout analysis models; the model trained without synthetic data tends to over-segment single lines in multiple segments, while the model trained with synthetic data is better at identifying curved or damaged single lines

Further improvements can be done on three dimensions: data, architecture and evaluation, and in different directions depending on the focus (lines or areas; realistic looking output for humans, or simply usable output for data augmentation tasks). Regarding data, improvements could be obtained by using datasets (or improving existing ones) to include semantic areas whose delimitation more closely matches eventual irregularities, overlaps, etc., of the actual images. Regarding the architecture, and its inputs, we could try to improve the line-based results by giving semantic types to lines based on the type of the zone in which they are contained (for the datasets that have this kind of information).

Finally, regarding the evaluation, due to the limits of native metric for assessing GAN, we propose to proceed in two complementary directions: machine-based and human-based. Machine-based evaluation will be perform by using generated pages as data augmentation for layout-oriented tasks, such as layout recognition, the rationale being that, if the inclusion of synthetic data leads to substantial score improvements, it can be inferred that relevant (and quantifiable) information is contained in the generation. This could be particularly relevant in the case of more complex layouts and small datasets.

Secondly, the qualitative assessment could be performed using a standardised protocol, e.g., historical and fake pages are presented to expert users through an application, in a controlled setting, and the user is asked to assess the nature of the page. The more the assessment of a single page deviates from the average score of a user in assessing the nature of pages (i.e., the more or less the user is "fooled"), the better or worse its score.

These dual evaluations will likely be complementary, as the information used by the machine or examined by expert users might not be the same. For instance, for layout-oriented automated tasks, it is not sure that the precise nature of the text or the quality of the decoration will weigh as much as they do for the expert. For this reason, further improvements could be performed as to maximise one or the other of these scores, or in both directions. The choice between one direction or the other could be determined by the final goal of the generation, i.e. concentrating on data-augmentation, or on the generation of historically plausible pages for human experts, for instance in the perspective of virtual restoration of damaged manuscripts (such as has been recently performed on a text-level for Greek inscriptions [2]).

To conclude, our method demonstrates remarkable versatility in generating convincingly authentic historical data through a streamlined protocol. We noted a general improvement in layout analysis tasks, particularly for resources that are typically underrepresented, thus broadening the scope for studies on less-explored HTR sources. Additionally, the simple yet effective creation of character masks paves the way for generating new texts from previously generated texts, potentially using a language model like an LLM. More broadly, the generated images, whether they depict real text (character mask) or simulated text (line mask), raise intriguing paleographic questions about the composition and perception of features relative to different scripts.

Acknowledgments. This research was funded by the Paris Artificial Intelligence Research Institute (PRAIRIE), under the Human and Social Sciences call, project "Artificial Pasts: Lost Texts and Manuscripts that never were" (P.I. J.B. Camps).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arroyo, D.M., Postels, J., Tombari, F.: Variational transformer networks for layout generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13642–13652 (2021)
2. Assael, Y., Sommerschield, T., Shillingford, B., Bordbar, M., Pavlopoulos, J., Chatzipanagiotou, M., Androutsopoulos, I., Prag, J., de Freitas, N.: Restoring and attributing ancient texts using deep neural networks. Nature **603**(7900), 280–283 (2022)
3. Barrere, K., Soullard, Y., Lemaitre, A., Coüasnon, B.: Training transformer architectures on few annotated data: an application to historical handwritten text recognition. International Journal on Document Analysis and Recognition (IJDAR) pp. 1–14 (2024)
4. Binmakhashen, G.M., Mahmoud, S.A.: Document Layout Analysis: A Comprehensive Survey. ACM Computing Surveys **52**(6), 109:1–109:36 (Oct 2019). https://doi.org/10.1145/3355610, https://doi.org/10.1145/3355610
5. Biswas, S., Banerjee, A., Lladós, J., Pal, U.: DocSegTr: an instance-level end-to-end document image segmentation transformer. arXiv preprint **arXiv:2201.11438** (2022)
6. Biswas, S., Riba, P., Lladós, J., Pal, U.: Docsynth: a layout guided approach for controllable document image synthesis. In: ICDAR 2021 – 16th International Conference on Document Analysis and Recognition. pp. 555–568. Springer (2021)
7. Clérice, T.: Ground-truth free evaluation of HTR on old French and Latin medieval literary manuscripts. In: Computational Humanities Research Conference (CHR) 2022 (2022)
8. Clérice, T.: You Actually Look Twice At it (YALTAi): using an object detection approach instead of region segmentation within the Kraken engine. Journal of Data Mining & Digital Humanities (2023)
9. Diem, M., Kleber, F., Fiel, S., Grüning, T., Gatos, B.: cBAD: ICDAR2017 Competition on Baseline Detection. In: ICDAR 2017 – 14th International Conference on Document Analysis and Recognition. vol. 01, pp. 1355–1360 (2017). https://doi.org/10.1109/ICDAR.2017.222
10. Diem, M., Kleber, F., Sablatnig, R., Gatos, B.: cBAD: ICDAR2019 Competition on Baseline Detection. In: ICDAR 2019 – 15th International Conference on Document Analysis and Recognition. pp. 1494–1498 (2019). https://doi.org/10.1109/ICDAR.2019.00240
11. Fogel, S., Averbuch-Elor, H., Cohen, S., Mazor, S., Litman, R.: Scrabblegan: Semi-supervised varying length handwritten text generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4324–4333 (2020)
12. Gabay, S., Camps, J.B., Pinche, A., Jahan, C.: SegmOnto: common vocabulary and practices for analysing the layout of manuscripts (and more). In: 1st International Workshop on Computational Paleography (IWCP@ ICDAR 2021) (2021)

13. Grüning, T., Labahn, R., Diem, M., Kleber, F., Fiel, S.: Read-bad: A new dataset and evaluation scheme for baseline detection in archival documents. In: 2018 13th IAPR International Workshop on Document Analysis Systems (DAS). pp. 351–356. IEEE (2018)
14. Grüning, T., Leifert, G., Strauß, T., Michael, J., Labahn, R.: A two-stage method for text line detection in historical documents. International Journal on Document Analysis and Recognition (IJDAR) **22**(3), 285–302 (2019)
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
16. Hoyez, H., Schockaert, C., Rambach, J., Mirbach, B., Stricker, D.: Unsupervised Image-to-Image Translation: A Review. Sensors **22**(21) (2022). https://doi.org/10.3390/s22218540, https://www.mdpi.com/1424-8220/22/21/8540
17. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-Image Translation with Conditional Adversarial Networks. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on (2017)
18. Kahle, P., Colutto, S., Hackl, G., Mühlberger, G.: Transkribus-a service platform for transcription, recognition and retrieval of historical documents. In: ICDAR 2017 – 14th International Conference on Document Analysis and Recognition. vol. 4, pp. 19–24. IEEE (2017)
19. Kiessling, B., Ezra, D.S.B., Miller, M.T.: BADAM: a public dataset for baseline detection in Arabic-script manuscripts. In: Proceedings of the 5th International Workshop on Historical Document Imaging and Processing. pp. 13–18 (2019)
20. Kiessling, B., Tissot, R., Stokes, P., Ezra, D.S.B.: eScriptorium: an open source platform for historical document analysis. In: ICDAR 2019 – 15th International Conference on Document Analysis and Recognition, Workshops (ICDARW). vol. 2, pp. 19–19. IEEE (2019)
21. Madi, B., Alaasam, R., Shammas, R., El-Sana, J.: Scheme for palimpsests reconstruction using synthesized dataset. International Journal on Document Analysis and Recognition (IJDAR) **26**(3), 211–222 (2023)
22. Monnier, T., Aubry, M.: docExtractor: An off-the-shelf historical document element extraction. In: 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 91–96. IEEE (2020)
23. Muehlberger, G., Hackl, G.: NewsEye / READ OCR training dataset from French Newspapers (18th, 19th, early 20th C.) (Nov 2020). https://doi.org/10.5281/zenodo.4293602
24. Najem-Meyer, S., Romanello, M.: Page layout analysis of text-heavy historical documents: a comparison of textual and visual approaches. In: Proceedings of the Computational Humanities Research Conference 2022 Antwerp, Belgium, December 12-14, 2022. pp. 36–54 (2022)
25. Nikolaidou, K., Seuret, M., Mokayed, H., Liwicki, M.: A survey of historical document image datasets. International Journal on Document Analysis and Recognition (IJDAR) **25**(4), 305–338 (2022)
26. Oliveira, S.A., Seguin, B., Kaplan, F.: dhSegment: A generic deep-learning approach for document segmentation. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 7–12. IEEE (2018)
27. Pang, Y., Lin, J., Qin, T., Chen, Z.: Image-to-image translation: Methods and applications. IEEE Transactions on Multimedia **24**, 3859–3881 (2021)
28. Pfitzmann, B., Auer, C., Dolfi, M., Nassar, A.S., Staar, P.: DocLayNet: a large human-annotated dataset for document-layout segmentation. In: Proceedings of

the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 3743–3751 (2022)
29. Pinche, A.: Cremma Medieval (June 2022), https://github.com/HTR-United/cremma-medieval
30. Pisaneschi, L., Gemelli, A., Marinai, S.: Automatic generation of scientific papers for data augmentation in document layout analysis. Pattern Recognition Letters **167**, 38–44 (2023). https://doi.org/https://doi.org/10.1016/j.patrec.2023.01.018, https://www.sciencedirect.com/science/article/pii/S0167865523000247
31. Poddar, A., Dey, S., Jawanpuria, P., Mukhopadhyay, J., Kumar Biswas, P.: TBM-GAN: Synthetic Document Generation with Degraded Background. In: International Conference on Document Analysis and Recognition. pp. 366–383. Springer (2023)
32. Quirós, L.: Multi-task handwritten document layout analysis. arXiv preprint **arXiv:1806.08852** (2018)
33. de Sousa Neto, A.F., Bezerra, B.L.D., de Moura, G.C.D., Toselli, A.H.: Data Augmentation for Offline Handwritten Text Recognition: A Systematic Literature Review. SN Computer Science **5**(2), 258 (2024)
34. Stoekl Ben Ezra, D., Brown-DeVost, B., Jablonski, P., Lapin, H., Kiessling, B., Lolli, E.: BiblIA - a General Model for Medieval Hebrew Manuscripts and an Open Annotated Dataset. In: The 6th International Workshop on Historical Document Imaging and Processing. p. 61–66. HIP '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3476887.3476896
35. Tanveer, N., Ul-Hasan, A., Shafait, F.: Diffusion Models for Document Image Generation. In: International Conference on Document Analysis and Recognition. pp. 438–453. Springer (2023)
36. Vidal-Gorène, C., Camps, J.B., Clérice, T.: Synthetic lines from historical manuscripts: an experiment using GAN and style transfer. In: International Conference on Image Analysis and Processing. pp. 477–488. Springer (2023)
37. Vidal-Gorène, C., Dupin, B., Decours-Perez, A., Riccioli, T.: A modular and automated annotation platform for handwritings: evaluation on under-resourced languages. In: ICDAR 2021 – 16th International Conference on Document Analysis and Recognition. pp. 507–522. Springer (2021)
38. Vidal-Gorène, C., Lucas, N., Salah, C., Decours-Perez, A., Dupin, B.: RASAM – A Dataset for the Recognition and Analysis of Scripts in Arabic Maghrebi. In: Barney Smith, E.H., Pal, U. (eds.) ICDAR 2021 – 16th International Conference on Document Analysis and Recognition, Workshops (ICDARW). pp. 265–281. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-86198-8_19
39. Vögtlin, L., Drazyk, M., Pondenkandath, V., Alberti, M., Ingold, R.: Generating synthetic handwritten historical documents with OCR constrained GANs. In: ICDAR 2021 – 16th International Conference on Document Analysis and Recognition. pp. 610–625. Springer (2021)
40. Wang, H., Wang, Y., Wei, H.: Affganwriting: a handwriting image generation method based on multi-feature fusion. In: International Conference on Document Analysis and Recognition. pp. 302–312. Springer (2023)
41. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: Layoutlm: Pre-training of text and layout for document image understanding. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 1192–1200 (2020)

42. Zhong, X., Tang, J., Yepes, A.J.: Publaynet: largest dataset ever for document layout analysis. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1015–1022. IEEE (2019)