Artificial colorization of digitized microfilms and its impact on other tasks

Thibault Clérice, Ariane Pinche

Journal of Data Mining and Digital Humanities

Artificial colorization of digitized microfilms and its impact on other tasks

Thibault Clérice[1,2] **and Ariane Pinche**[1,3]

[1]Centre Jean Mabillon, École des chartes – PSL, Paris, France
[2]UMR 5189, HIstoire et SOurces des Mondes Antiques, Université de Lyon, France
[3]UMR 5648, Histoire, Archéologie, Littératures des mondes chrétiens et musulmans médiévaux, France

Corresponding author: Thibault Clérice , thibault.clerice@chartes.psl.eu

Abstract
A lot of available digitized manuscripts online are actually digitized microfilms, a technology dating back from the 1930s. With the progress of artificial colorization, we make the hypothesis that microfilms could be colored with these recent technologies, testing *InstColorization*. We train a model over an *ad-hoc* dataset of 18 788 color images that are artificially gray-scaled for this purpose. With promising results in terms of colorization but clear limitations due to the difference between artificially grayscaled images and "naturaly" greyscaled microfilms, we evaluate the impact of this artificial colorization on two downstream tasks using *Kraken*: layout analysis and text recognition. Unfortunately, the results show little to no improvements which limits the interest of artificial colorization on manuscripts in the computer vision domain.

> Many low resolution digital scans of microfilms exist. These are surrogates of surrogates. They can still be (and are) profitably used, for example to corroborate a particular reading. I am however skeptical of using them as a single source for making an edition. Perhaps, indeed, 99% of a manuscript can still be deciphered by using them, but it is about that 1% of cases in which the scribe fumbled a bit with his pen and it is unclear what the word reads. In those 1% cases, you do not wish to have a low-resolution, black and white reproduction of a reproduction as your sole witness
>
> L. W. C. van Lit Lit [2019]

I INTRODUCTION

In the digital age, and specifically in the last couple of years with the pandemic, transcribing using digital surrogates has become a common way to work through the edition of medieval works. However, if the technology of high definition and color digital photography is much preferred for these goals, we still often encounter microfilms of ancient manuscripts. In the late 1920s and early 1930s (Nicoll [1953]), the microfilm technology allowed for the constitution of early collections of manuscripts' surrogates. These collections, first aimed at facilitating the access of remote manuscripts (Stevens [1950]), was also pursued for other reasons such as preserving manuscripts in the context of the Second World War (Project [1944]). In France, the "Institut de Recherche en Histoire des Textes" (IRHT) had started photographing manuscripts first, moved to microfilming until digital photography's quality became good enough to switch technology (Holtz [2000]). With an abundance of microfilms, while transitioning to color capture, cultural heritage institutions have made digitized microfilms available to the wide audience through platforms such as Gallica at the bibliothèque nationale de France.

Indeed, scanning microfilm implies lower risks[1] and as such lower costs that it is a quicker way to build a first online collection. As a matter of fact, in 2001, the cost of microfilm digitization was ten times lower than the one of color digitization (Council on Library & Information Resources [2001]). As of 2021, while microfilming scans are generally clearly priced, such as the pricing at the New York Public Library (150$ per reels), digitization of old, fragile and rarer books are mostly quoted on demand, or rated per day or hour of work (350$ at NYPL)[2]. Until cultural heritage institutions have the ability and time to digitize all their manuscripts in color, for many works, the only available version remains digitized microfilms. But there are situations where this will not be possible: unfortunately, there are manuscripts whose only surviving exemplars are microfilm surrogates, such as many manuscripts from the *bibliothèque municipale de Chartres* whose microfilms were produced by the IRHT before they were destroyed during WW2[3]. For these manuscripts, their original colors are lost forever.

Nowadays, with the recent development of deep learning colorization, there is hope for the ability to bring back some color to these old microfilms. Colorizing deep learning networks using generative adversarial networks (GAN) or other architectures showed the ability for colorizing so-called black and white photographs. Colorizing more than 100-year-old pictures of ancestors and villages has been widely popular in news media and social ones - enough for companies to create business around it[4] -, and it provides an interesting testing ground for computer vision. The computer science side of this work has heavily outweighted the humanities and social sciences side in this domain: most if not all papers on this topic follow the same dataset-new model-output based architecture with little to no space for questioning the results[5]. And if this technology of colorization seems to be of more interest for outreaching strategies of cultural heritage institutions than for researchers, the bias behind these colorization has been clearly shown, with ethical questions raised by many: people tend to be whitened in specific contexts (Goree [2021]), colorful local dresses made grayish or dulled (Katz [2021]). In this context, using these colorization to study the past would raise epistemological questions more than they would provide insight. In the context of manuscripts though, if the output cannot be taken as the real colors that X manuscript had, this colorization might also provide a better and clearer image of manuscripts, by automatically balancing colors and providing more than one color space and enhance the efficiency of downstream tasks for text acquisition. Of course, this can also be of interest for libraries and the likes in terms of outreaching, while disclaiming the same limitations as the one for photographs.

In this context, we put to test a deep learning colorization tool - *InstColorization* from Su et al. [2020] - on digitized manuscripts. In order to do so, after a clear description of the experimental setup including brand new datasets for colorization, manuscript segmentation and handwritten text recognition (HTR). Not only do we investigate the possibilities of colorization

[1]Destroying or damaging a manuscript would lead to great loss in terms of knowledge, while harming microfilms most often simply represents a delay for the presence of the original manuscript online.

[2]https://www.nypl.org/help/get-what-you-need/bookscanning

[3]On 26 May 1944, the town hall of Chartres containing the library was bombed by the American air force. “45% of the manuscripts were totally destroyed”, while the remaining survived in varying states (https://www.manuscrits-de-chartres.fr/fr/incendie-et-ses-consequences)

[4]See the services of MyHeritage for example, https://www.myheritage.fr/incolor, which states in French "Import dull black and white picture and be *fascinated* by the results

[5]See Joshi et al. [2020], Boutarfass and Besserer [2020], Chen et al. [2018]. We must note that for the later, the question of dataset bias was taken into account at least for the purpose of colorizing a specific and “new” domain, Chinese black and white films

for interacting with microfilms, but we also evaluate the potential impact of this post-processing step for both layout segmentation and HTR. Finally, we propose a road-map for stronger evaluation of the impact of colorizing microfilms on readers while discussing the current prowess and limitations of the output.

II METHOD

We design an experiment that aims at both producing colorized microfilms and evaluating the impact of colorization on other computer vision tasks. We train a model for colorization using an *ad-hoc* corpus for the paper which we evaluate qualitatively. Then, we train Handwritten Text Recognition models alongside layout segmentation ones with *Kraken* which we attempt to evaluate quantitatively.

2.1 Colorization

2.1.1 Dataset

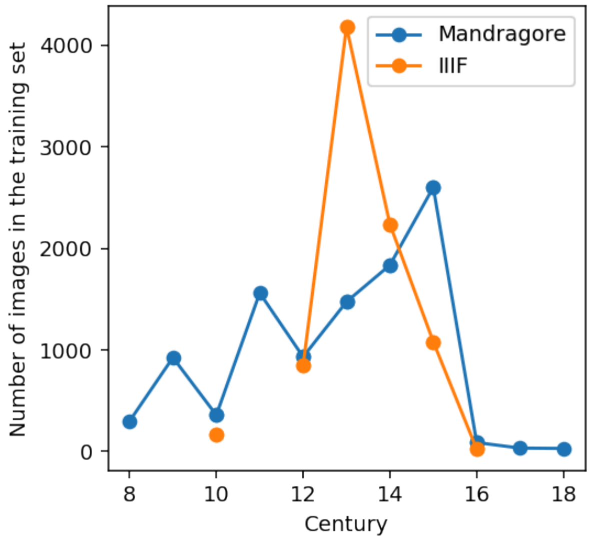

Figure 1: Distribution of pictures per century. IIIF pictures are drawn from manuscripts and mostly come with other folios. On the other hand, Mandragore data are more likely to be taken from different original materials.

In order to train the colorizer, we offer a dataset that is built around 5 cornerstones:

- It is chronologically centered around the Middle Ages, from the 8th century to the 16th (see Figure 1).
- It is focused on west European manuscripts.
- It is balanced in accordance with overall numbers of surviving manuscripts while not allowing a standard deviation of the population per century too high, allowing the resulting models to be generic and fine-tuned if necessary.
- It must contain both highly decorated pages and very simple ones.

This results in an 18 788 files-large dataset made from two data sources. 8 660 digitized pages or cover in color from 50 different manuscripts mostly coming from the *Gallica* and *e-Codices* platforms, which might or might not display decorated elements, illustrations and texts (hereafter conveniently named as *IIIF* in figures). In this part of the dataset, 3 "manuscripts" were specifically chosen for the diversity of their content over time as they were composite books made of manuscript fragments and display overlap of papers (see the second row of

Figure 2). Some of these also marginally contain printed content. For the second part of the corpus, we used the *Mandragore* database (Aniel [1992]), a dataset whose aim is to collect and annotate illuminations and decorations in general in manuscripts. This part is composed by 10 128 pictures randomly drawn from the occidental manuscripts[6], distributed over 2 612 different original codices. Those pictures can be complete or cropped page centered around a specific illustration. 1 903 manuscripts of the 2 612 are represented by a single picture, 23 manuscripts have more than 50, 3 have more than 100. The overall dataset presents paper colors (including shades), black ink for the text; as far as colors go, red and blue are dominant in drop capitals, decorations, illustration and rubricated texts, green and gold are present in the dataset in a smaller fashion.

Figure 2: First row: random examples of pictures from the Mandragore part of the training dataset. Second row: 2 pages from composite books in the IIIF part of the dataset and 3 random one from "original" manuscripts. Images display ratio are changed for the purpose of displaying multiple example on the same page.

Thanks to the dedication towards open-access and open standards of cultural heritage institutions such as the bibliothèque nationale de France and projects such as "e-Codices", we were able to download and produce this rather large dataset (14.8 GB) for a first experiment: it must be stressed how much the availability of data on their platform can and do help computational research in building large datasets.

2.1.2 Training set-up

As our aim is not to develop a colorizing network but rather to evaluate for the first time its capacities toward microfilms, the experimental set-up requires 3 capacities from colorizing tools: it should be well described by a paper, it must be possible to install it with clear dependencies, it must provide a predicting feature that over-goes the classical test score production. Given these predicaments, we reuse the tool from "Instance-aware Image Colorization" (Su et al. [2020]), thereafter *InstColorization*, which provides a simple, detailed and reusable way to train and fine-tune models. *InstColorization*[7] is built around 3 colorization networks, each trained independently, as well as an object detection one (see Figure 3) which is not directly trainable with the available code in the repository but reuses an external pre-trained parameter file.

The training is run with the same parameters from the original scripts of the paper with half precision to reduce training time: the final training time takes a little under 78 hours with

[6]Randomness only affects the selection inside group of centuries.

[7]https://github.com/ericsujw/InstColorization

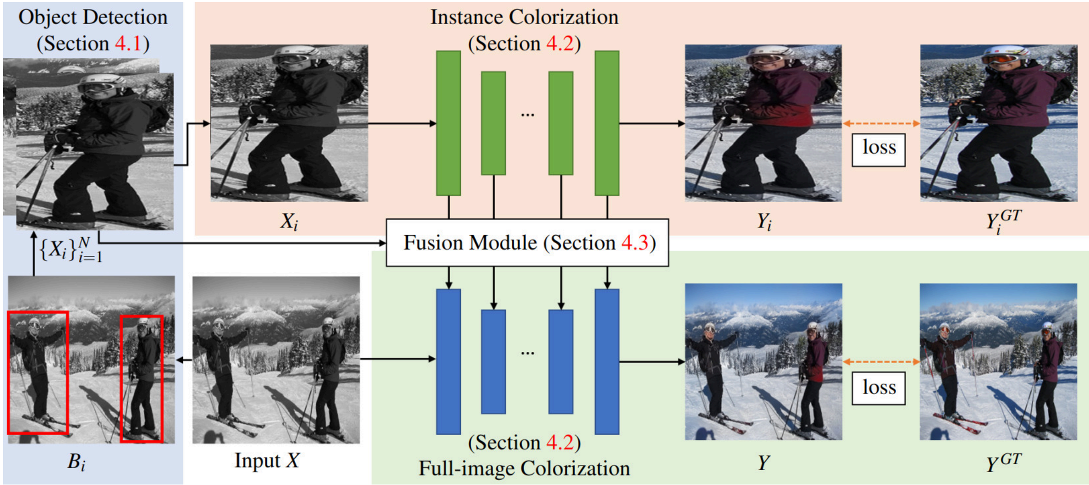

Figure 3: Method overview drawn from "Instance-aware Image Colorization" (Su et al. [2020]). An image is colorized entirely (4.2, bottom) while each detected object in it (4.1) is also colorized independently (4.2 up). The final result is the output of a fusion networks (4.3), which takes the output of previously colorized objects and the full-image in order to produce a general colorized output.

a RTX2080TI GPU spanned over 150 epochs for both the full and detected object instances and 30 epochs for the fusion model with 256x256 pixels resized input. The network itself is responsible for grayscaling images and we do not tweak the software outside of its way of loading images and their detected objects[8]. The final dataset presented earlier is actually the result of a filtering by the model when the object detection model did not yield any result for detected objects: the first version of our dataset was as large as 27 415 files, but 8 620 of its members were rejected by the object detection method. This probably advocates for training object detection on such materials in the future if the road that is taken by *InstColorization* is validated as a state-of-the-art method on the long run.

2.2 Text acquisition tasks

2.2.1 Dataset

For all computer vision tasks, our dataset is based on the CREMMA Medieval dataset (Pinche [2021a]), built with eScriptorium (Kiessling et al. [2019]), an interface for HTR ground truth production, and *Kraken* (Kiessling [2019]), an HTR and layout segmentation engine. It is composed of seven Old French manuscripts written between the 13th and 14th centuries (see Table 1), mainly scanned in high definition and color except for one manuscript (Vatican) which is a black and white document, most probably binarized by the holding library. As the dataset is made from pre-existing transcriptions, the sample size is very different from one source manuscript to the other. The basis of the dataset is composed of a graphemic transcription (see after for the transcription principles) of *La Vie de saint Martin* and of the *Dialogues sur les vertus de saint Martin* from the hagiographic collection *Li Seint Confessor* of Wauchier de Denains (Pinche [2021b]). It is supplemented with carefully aligned data from other projects: transcriptions[9] of *Chanson d'Otinel* (Camps [2017]) from the Geste project[10],

[8] We simply made sure that files containing dots such as `x.y.jpg` would not be collapsed into single files with the original implementation looking for dot as separators between file extension and filenames (`x.jpg`)

[9] Cologny, Bodmer, 168 and Vatican, Reg. Lat., 1616

[10] https://github.com/Jean-Baptiste-Camps/Geste

transcription of *Manuscrit du Roi* for the Maritem project[11], crowdsourced transcriptions of the collaborative projects of the Standford Library: Image du Monde (BnF. Bibliothèque de l'Arsenal. Ms-3516[12]) and the Bestiaire de Guillaume le Clerc de Normandie (Bibliothèque nationale de France fr. 2442[13]), and a few folios transcribed in relation with the new project of editing *Les Sept Sages de Thèbes*[14].

As the data come from different projects, it is standardized to strengthen any layout segmentation and HTR model[15]. The layout segmentation follows the Segmento ontology[16], separating the main column, margin, numbering, drop capital. We chose a graphemic[17] transcription method to have a sign in the image corresponding to a sign in our text: all the abbreviations are kept, and *u/v* or *i/j* are not dissimilated. But we exclude graphetic transcription method which distinguish different forms of letters (*e.g.*, *s* and *long s*), considering that it would reduce the number of examples per character, as well as highly impact digital transcription time and introduce even more disagreement between transcribers[18]. Finally, the spaces in the dataset are not homogeneously represented in the ground truth text annotation, with transcriber reproducing the manuscript spacing while others use lexical spaces. It must be stressed that spaces are the most important source of error in medieval HTR models[19]: for the model Bicerin (Pinche [2021a]), spaces represent 33.9% of errors[20]). In the current state of the art of HTR, some workflows (Camps et al. [2021, 2020]) chose to solve this problem with a secondary tool such as Boudams (Clérice [2019]), a deep learning tool built for word segmentation in Latin or Medieval French.

Manuscript Identifier	Date	Pages	Columns	Lines	Color
BnF, Arsenal 3516	13th	10	40	1991	Yes
BnF, ms fr. 22549	14th	3	9	411	Yes
BnF, ms fr. 24428	13th	20	40	1295	Yes
BnF, ms fr. 412	13th	49	98	4551	Yes
BnF, ms fr. 844	13th	18	36	1026	Yes
Cologny, bodmer, 168	13th	22	44	1927	Yes
Vaticane, Reg. Lat., 1616	14th	41	41	1726	No
Total		163	308	12927	

Table 1: CREMMA Medieval dataset statistics. Sample pictures are available in the appendix, Figure 13.

Of these, the microfilmed manuscripts (see Table 2) all dating from the end of the 13th century or the 14th century and written in Old French are kept for evaluating performances as our test dataset[21]. They all come from hagiographic collections and present dialectal differences. In

[11]Produced by V. Mariotti within https://anr.fr/Projet-ANR-18-CE27-0016

[12]https://fromthepage.com/stanfordlibraries/image-du-monde-en-vers

[13]https://fromthepage.com/stanfordlibraries/guillaume-le-clerc-de-normandie-s-bestiary

[14]Manuscript, BnF, fr.22550, this project just started in Geneva under the direction of Y. Foehr (Geneva) and S. Ventura (Brussels).

[15]To ensure the quality of the data, continuous integration workflow were put in place checking the segmentation vocabulary (HTRVX, Clérice and Pinche [2021b], a XML schema validator), but also the homogeneity of the signs of the characters used in the dataset through a list of authorized signs and translation table with ChocoMufin, Clérice and Pinche [2021a]

[16]https://github.com/SegmOnto/examples

[17]We use the terminology graphemic (*graphématique*) and graphetic (*allographétique*) following D. Stutzmann definitions, see Stuzmann [2011]

[18]Distinction of specific forms can be difficult and would require only expert transcribers for such corpora

[19]And it definitely happens that editors disagree between agglutinated and split forms of words.

[20]The same issue was found on Transkribus and a previous version of Kraken two years before (Camps et al. [2020]

[21]All the microfilm scans come from Gallica and all the original manuscripts come from French manuscripts department of the bibliothèque nationale de France (BnF).

all these sources, the layout is similar and the writing is easy to read with rare but present abbreviations. The text is organized in 2 columns per page and in 40 or 42 written lines if there is no decoration. The scanned microfilms each present two pages. Each of them has some peculiarities: marginalia, interlinear lines, running titles, rubrics, drop capitals or decorations that can make segmentation difficult. In our case, the manuscript fr. 411, an unfinished manuscript, presents the case of a missing capital and decoration and shows instead white spaces waiting to be illuminated. The segmentation test dataset only differ on its size, as each manuscript contains more documents, ranging from 3 to 6 documents, equivalent to 6 or 12 pages.

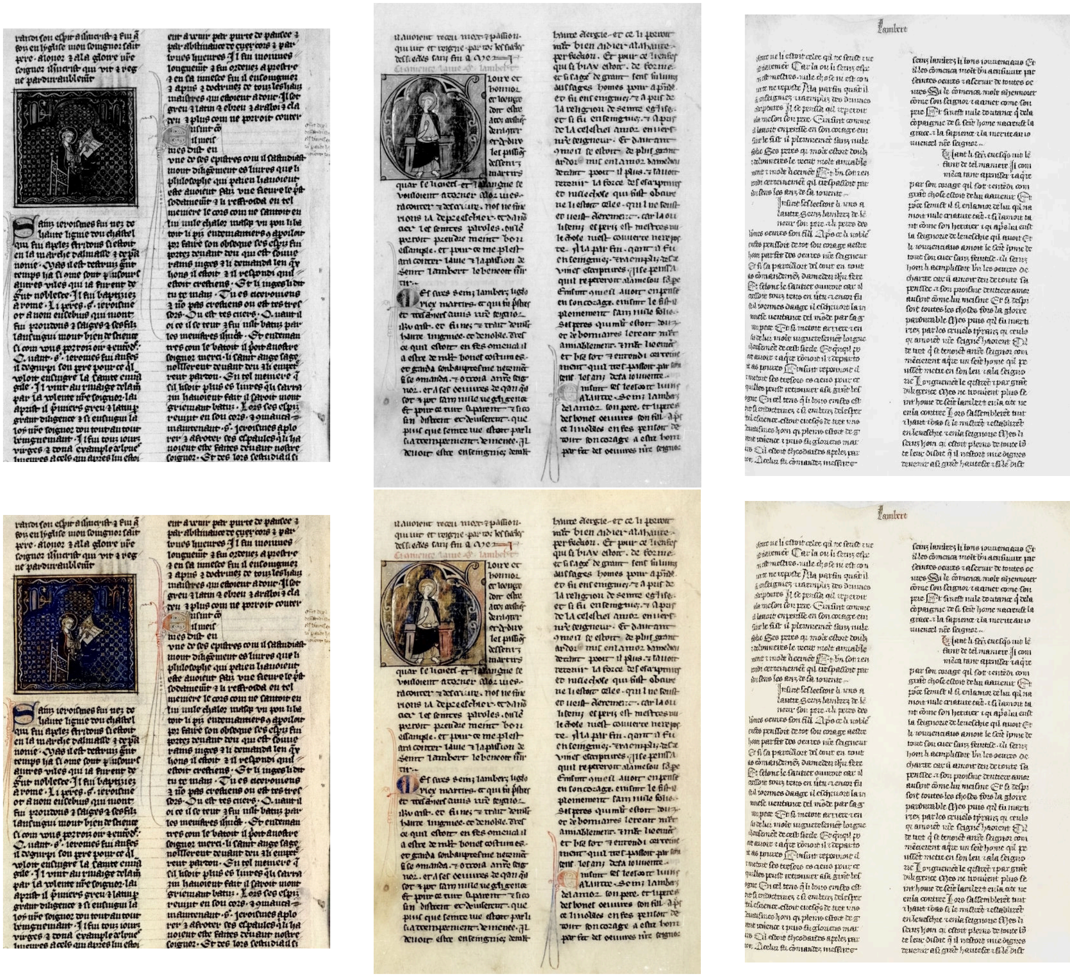

Figure 4: Partial reproduction of manuscripts BnF, ms. fr. 13496, 17229 and 411 (a single page is shown for convenience.) Second row contains the colorized version. While the colorization is not perfect, we see that some patterns are recognized, namely the blue background on fr. 13496 and the dull but colorized drop capitals on the first two manuscripts.

Manuscript ID Date Annotated text	Ms, fr. 13496 Late 13th c. Saint Jerome's Life	Ms, fr. 17229 Late 13th c.– early 14th c. Saint Lambert's Life	Ms, fr. 411 14th c. Saint Lambert's Life
HTR Ground Truth Locus	fol. 245v - fol.246r	fol. 163v - fol.164r	fol. 125v - fol.126r
No. of Document	1	1	1
No. of transcribed columns	4	4	4
No. of transcribed lines	159	160	150
Segmentation Ground Truth Locus	fol. 245v - fol.248r	fol. 163v - fol.169r	fol. 125v - fol.131r
No. of Document	3	6	6
No. of Segmented Columns	12	24	24
Decoration	yes	yes	no
Drop Capital	yes	yes	no
Rubric	no	yes	no
Running Title	no	no	yes
Numbering	yes	yes	yes
Marginalia / interlinear	yes	no	no

Table 2: Presentation of the digitized microfilmed manuscripts. Documents are double paged, composed by the verso of a folio and the recto of the following one.

2.2.2 *Training Set Up*

As *Kraken* is able to make use of color channel - since a couple year - at training and prediction time for both layout and HTR, avoiding any binarization or gray-levelling at the preprocessing step (Kiessling [2020]), we use *Kraken* for all our training and evaluation steps[22]. We train two different models for the segmenter: one for lines where they are grouped into a single category (numbering, rubricated and normal - *Default* in Segmonto - lines are merged into a single class), one for regions but only for main regions, as preliminary expriments show that illustrations and drop capital are not well recognized yet given the size of this dataset. We train models for 50 epochs and keep the best model of all accross all metrics.

For HTR, we train models over two versions of the same training and evaluating sets from the CREMMA project (microfilms excluded): in one version, we use all the available data (thereafter called BW models), in the second (NOBW models) we withdraw the black and white manuscript (Vatican, Reg. Lat., 1616). We run training procedures for HTR models with *Kraken* with and without augmentation for each version of the corpus (BW and NOBW). After training, we take the 10 best models of each training run, summing up to 40 models in total: each combination of augment and dataset version always reaches over 90% accuracy for its 10 best models on the evaluation set.

2.2.3 *Test Set Up*

Unfortunately, we were not able to produce a quantitative test for the region segmentation. Indeed, we could find only two existing tools for this purpose. The first one was built by PRIMA (Clausner et al. [2011]) and is working only with Windows, required a different version of PageXML from the one produced by *Kraken* as well as binarized images. The second one[23], built originally for ICDAR 2017, required images creation with various level of colors that are incompatible with some of our data, when we have overlap of zones (Alberti et al. [2017]). We fallback on qualitative evaluation of predictions, mostly looking at difference in between both models.

[22]The oldest public record we could find was the issue 117 from 2019 on Kraken's repository where Ben Kiessling provided us with the color configuration (https://github.com/mittagessen/kraken/issues/117)

[23]https://github.com/DIVA-DIA/DIVA_Layout_Analysis_Evaluator

Secondly, for baseline and mask evaluation, we predict segmentation with the produced model and evaluate it using the same tool[24] as ICDAR 2017 and 2019 for evaluating line segmentation (Alberti et al. [2017]), comparing segmentation results for both the original black and white microfilm and the colorized one.

Finally, we evaluate HTR performances' gains or drops by comparing the accuracy of models on the microfilm and colorized microfilms. Each of the 40 models is used to evaluate the prediction against the original ground truth, we then compare each test result from the colorized output with the microfilm original picture so that we retrieve a delta accuracy: $\Delta = \text{Accuracy}_{Colorized} - \text{Accuracy}_{Microfilm}$.

III RESULTS

3.1 Colorization

Figure 5: Output of training in epoch 145 (instance network), 149 (full network) and 29 (fusion network). The first column contains original colored images, the second automatically grayscaled ones, the last one the production from the network.

The training output is promising in the three different training phases (see Figure 5). On the training set, the paper of the manuscript is clearly distinguished from the background, illustrations are well colorized, including gold and green that are rarer overall, illustrated capitals (such as the E and 2 Qs in the second row of Figure 5) and rubricated texts are correctly colorized.

We then applied the models on to different microfilms (lost microfilms for Chartres, test microfilms for HTR and segmentation, random microfilms from *Gallica*) as well as artificial grayscaled color digitizations. First of all, the colorization algorithm works really well on grayscale images

[24]https://github.com/DIVA-DIA/DIVA_Line_Segmentation_Evaluator

(see Figure 6) and images without any decoration (see Figure 8): the page is correctly identified and colorized differently from any part of the background. In the context of photography artifacts present in the picture, such as what seems to be a clamp in Figure 8, the colorization might be hazardous while not affecting the overall colorization of the microfilm. For artificially grayscaled image, colors are retrieved correctly, including green and gold, and the colors do not seem to be less colorful than the original. Finally, colored inks on microfilms are variably predicted, ranging from hints of colors (see Figures 9 and 2) to probably correct but dulled colors (see Figure 8), rubricated text being the less recognized by the colorizer on our test set.

Regarding the limitations of artificial colorization of colored inks on microfilm, we make the hypothesis that the issue comes mostly from the difference in the contrast present in microfilms and artificially grayscaled images which form our training set. We confirm it by having a closer look at manuscripts such as BnF fr. 24369 whose content is both available in color and microfilms: it is clear that the contrast ratio and color levels are extremely different between the two. Given the age of microfilms, it is not impossible that, as it was the only possible surrogate at the time, contrasts were intensified at the time of photography in order to make the content easier to read rather than trying to capture the diversity of contrasts over a full folio[25]. We reproduce similar levels of contrast on artificially grayscaled images by adjusting color levels[26] with a page from one of the manuscripts of our training set (BnF, fr. 24369) and the output is clearly as bad if not worse than the output based on the digitized microfilm (see Figure 9). Regarding this dullness of colors and the difficulty for the resulting model to address this, we make the hypothesis that introducing Contrast and Color level transformations during training might improve the robustness of the prediction: we propose a first experiment of artificially corrected microfilms using CLAHE, Posterization and JitterColors from the *Albumentations* library that shows different levels of vibrant colors, sometimes at the expanse of more important bleed-through of the verso of each page (see Figure 10). These first results show the limit of post-processing as much as they advocate for in-training grayscale transformations to allow for better prediction on microfilms as the model learn to colorize images.

Regarding readability improvement for human readers, we are unable to propose a clear answer about the potential improvement produced by the current paper. For sure, as surrogates of surrogates, digitized microfilm often proposes a lower level of detail than their colorized counterparts and as such, when comparing high definition of color digitization with colored microfilms, the level of detail and readability is much higher in the first. However, for quantifying the impact of artificial colorization, we can only recommend to crowdsource the evaluation in further work, using the same approach as Manjavacas et al. [2019] while evaluating generated rap lyrics. The evaluation of such output should include

- Crowdsourced transcriptions of microfilms and colored microfilms followed by a quantitative evaluation of inter-annotator agreement. Artificially grayscaled and color scans must be part of the experiment as well to provide a baseline.
- Accompanying the first, on the same dataset, a readability scoring polls.
- Finally, an appreciation scoring polls, to quantify how much of interest colorization is.

[25] Also, the photographs were in black and white...

[26] We applied the following number in Gimp Color Level adjuster: Black 170, Clamp 4.25, White 255. The result is a bit darker at the border of the page, and the illustration is a little less readable

Figure 6: 13th century Latin Manuscript, artificially grayscaled digitization on the left, prediction on the right (BnF. Département des Manuscrits. français 375, 3v)

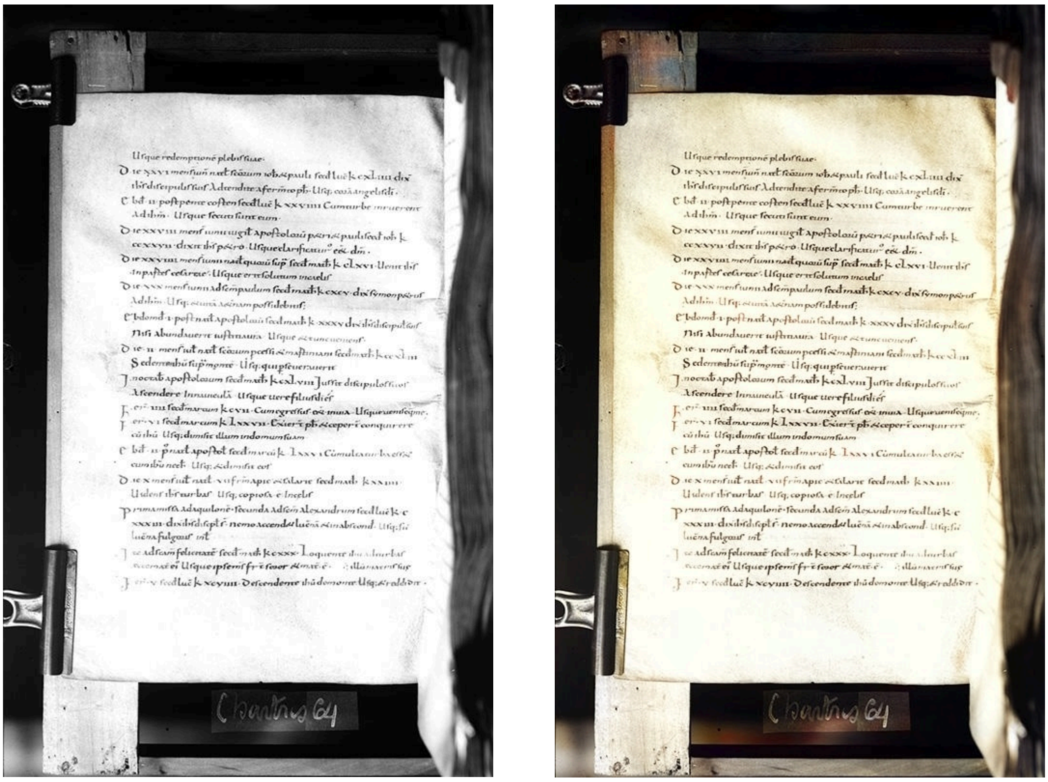

Figure 7: 9th century Latin Manuscript, lost in WWII, microfilm on the left, prediction on the right (Chartres, BM, ms. 64, unknown folio)

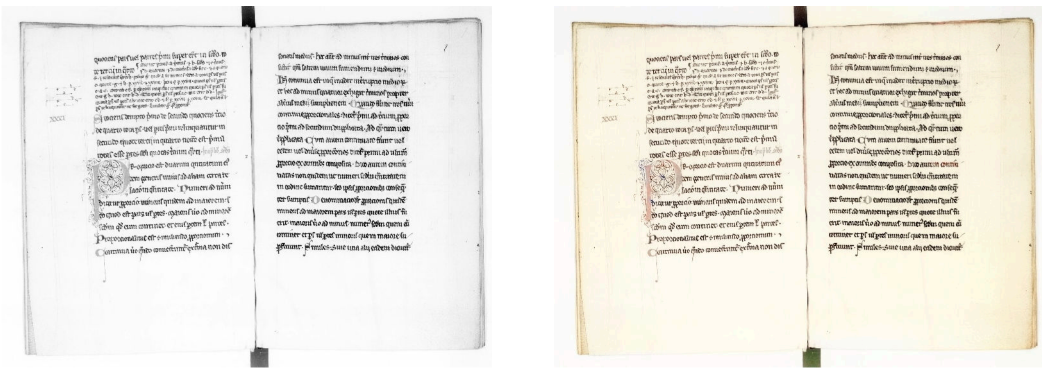

Figure 8: 13th century Latin Manuscript, microfilm on the left, prediction on the right (BnF. Département des Manuscrits. Latin 16644, 7v-8r)

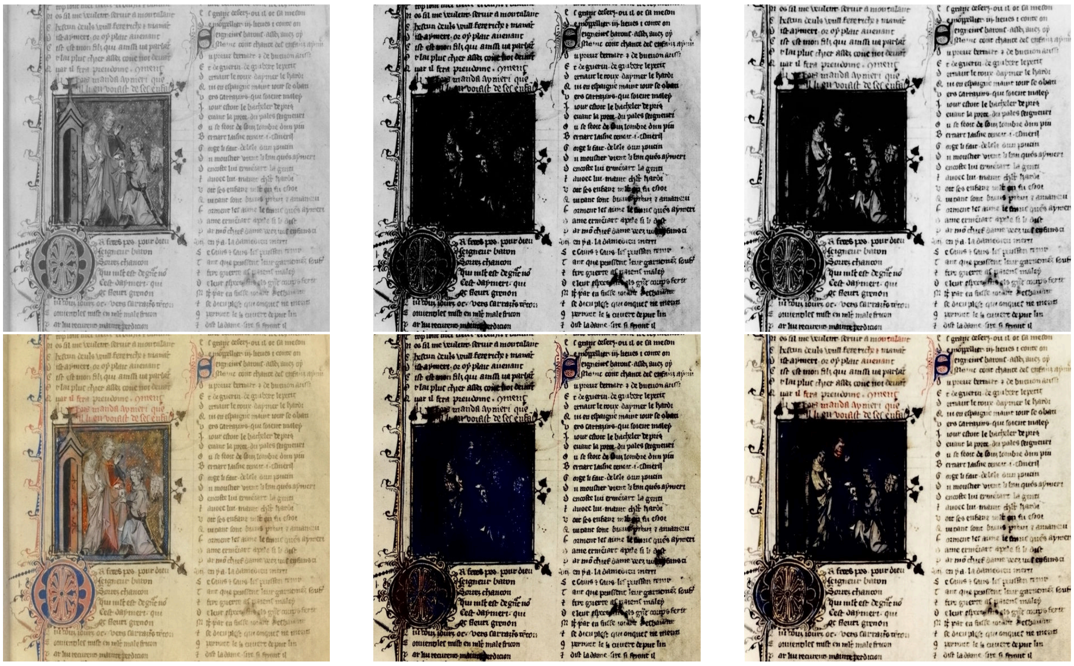

Figure 9: Examples with a section of manuscript whose original has been digitized by the BnF alongside its microfilm. First row contains grayscale images used as input for *InstColorization*, second row is the corresponding output. Columns are, in order: digitization of the manuscript in color, adjusted color level of the first column to come close to microfilm contrast, digitization of the microfilm. See Figure 14 for more examples.

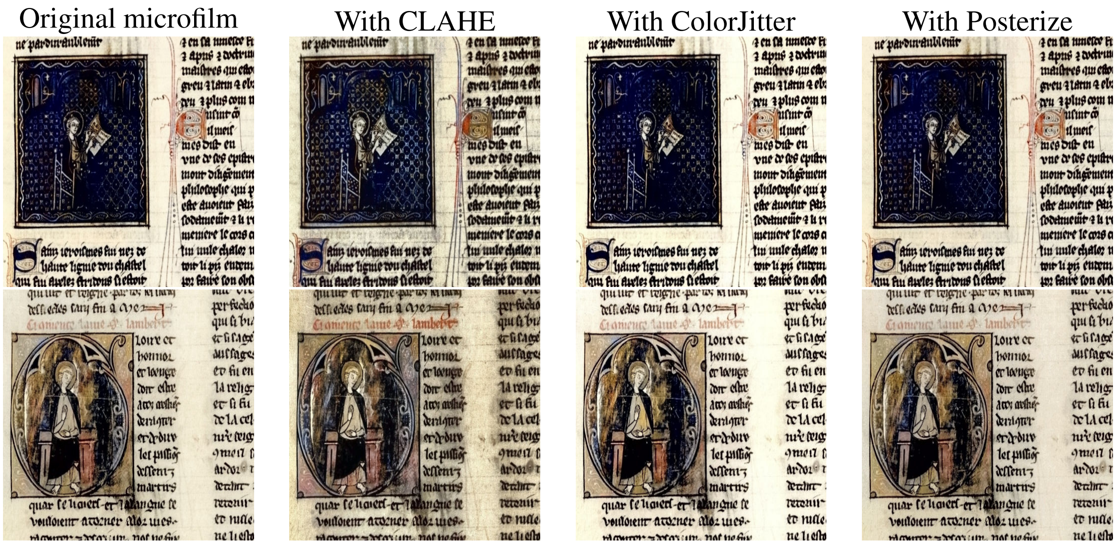

Figure 10: Impact on colorization of using pre-processing tools on original microfilms using the Albumentation libraryBuslaev et al. [2020] (details from fr. 13496 and ms. 17229)

3.2 Impact on computer vision tasks

3.2.1 *Layout segmentation*

Figure 11: Segmentation overlay of main regions on 125v-126r of fr. 411 and 163v-164r of fr. 17229. On the right is the original grayscale microfilm, on the left the artificially colorized one.

At the first level of segmentation, evaluated qualitatively, the impact of colorization is nearly null, with difference that is very small (see Figure 11 for examples). The region segmentation model has light difficulties around illustration on fr. 17229 - 163v with some light differences: the colored version is able to capture all the text from below the illustration while missing most of the rubricated text over it, while the original microfilm captured the later but missed a part of the text below. Both versions include some noise regarding the illustration [27].

At the second level, evaluated quantitatively, the impact is nearly null (less than 1% of improvement or metric drop) with a rather large variation (see Table 3). The manuscripts react again quite differently, specifically on the line metric, where fr. 13496 always profit from colorization while other manuscripts have much more nuanced results. It is clear that regarding segmentation, the model proposed by *Kraken* is resistent to color changes and much more centered on other features.

3.2.2 *Handwritten Text Recognition*

Finally, from the perspective of colorization, the mean accuracy delta over 120 tests (40 models $\times$ 3 manuscripts) is +0.025% while the median is 0.02% with the lowest outlier rating at -0.13% and highest one at +0.25% (see Figure 12). This means that the HTR models are as resistant to color changes as the segmentation models on *Kraken*. The delta varies depending on the manuscript (it seems that fr. 411 is affected more positively and more often than others: it might be due to the fact the manuscript is lacking illustrations and drop capitals) but so minimally that it should not advocate for colorizing manuscripts before running HTR pipeline. The use of *BW and NOBW* corpora and *Augmentation vs. No Augmentation* does not yield

[27] We provide the PageXML prediction output on our repository for each of the ground truth microfilm page for segmentation

Delta FMeasure for Manuscript	Lines	Matched Pixel	Pixel
fr. 13496, 245v-246r	0.00	-0.04	-0.03
fr. 13496, 246v-247r	0.00	0.02	0.02
fr. 13496, 247v-248r	0.32	0.00	0.00
fr. 17229, 163v-164r	0.31	0.02	0.25
fr. 17229, 164v-165r	0.31	0.00	0.01
fr. 17229, 166v-167r	0.31	-0.06	-0.06
fr. 17229, 167v-168r	-0.96	0.00	-0.45
fr. 17229, 168v-169r	0.35	-0.01	0.07
fr. 411, 125v-126r	0.32	0.11	0.15
fr. 411, 126v-127r	-0.31	0.08	0.07
fr. 411, 127v-128r	-0.61	0.03	0.00
fr. 411, 128v-129r	0.00	0.06	0.09
fr. 411, 129v-130r	0.31	0.04	0.29
fr. 411, 130v-131r	0.33	0.02	0.05

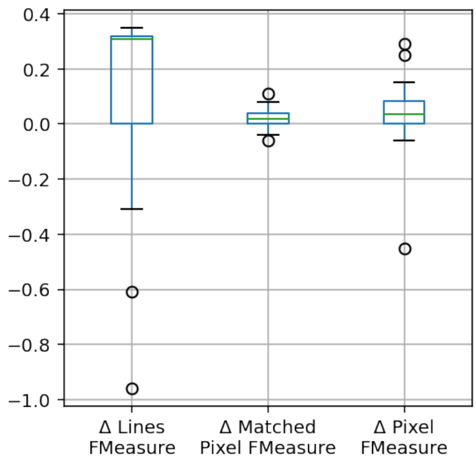

Table 3: Evaluating line segmentation impact: delta of different FMeasure between colorized and non-colorized of line detection. Positive outputs means colorized are better segmented than original microfilms. Metrics are in %.

any significant difference with a 0.04 median difference for the datasets and a null one for augmentation. From the perspective of HTR and *Kraken*, the results are however promising, as it shows a real capability of learning despite color differences, including when no grayscale or binarized images have been seen in the original training dataset.

IV CONCLUSION

Colorization of cultural heritage artifact through means of deep learning is a research area that has been left to the computer science side of research, while researchers in humanities and professional colorists are raising questions about the possible biased induced in this method. On the other hand, the general public has fancied the output of such tools to rediscover (fantasized ?) colors in their personal or local history. While photos and movies have been treated again and again, we made the hypothesis that colorizing manuscripts can be a new area of research, providing another version of manuscripts that are not yet digitized in color and as such helping readers, both human and computer, to better understand the content in the manuscript. We tested this hypothesis with *InstColorization* on a newly built dataset of nearly 20 000 manuscripts pictures from the western Medieval Ages. We are able to automatically color for the first time digitized microfilms: the background of pages are done efficiently while different inks are a little duller

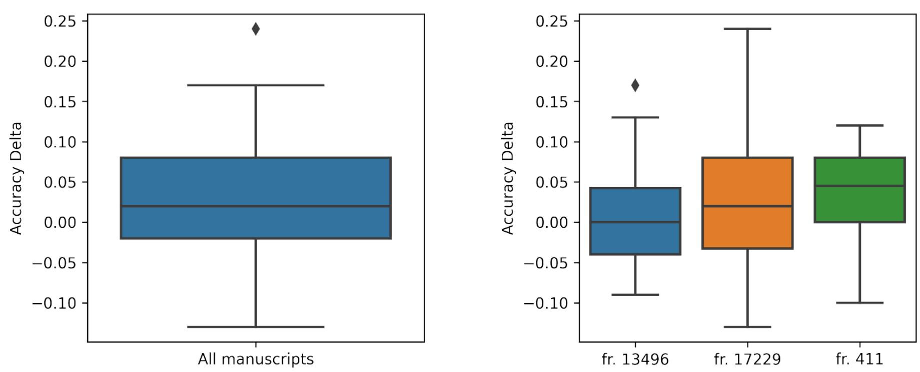

Figure 12: On the left, box plot of the accuracy delta with all 40 models regardless of the manuscript. On the right, box plot of the accuracy delta with all models distinguishing manuscripts.

than expected. We believe that these results can be enhanced by further works, namely working on the color-to-grayscale conversion to be able to deal with the different color curve present in microfilm digitization that the one we obtain with straight color-to-grayscale transformation or other tools that does more than adding channels without changing the contrast of the original picture. As for evaluation of the output from a human readability point of view, we only propose a way to evaluate such new versions of the manuscripts, which should be tested on a wide audience through transcription, cross- and auto-evaluation of participating readers. For computer vision downstream task such as layout segmentation and HTR, the colorization does not seem to affect at all the later and struggles to enhance the first.

V ACKNOWLEDGEMENTS

We thank Jean-Baptiste Camps and Simon Gabay for their insight and their help in choosing manuscripts for the IIIF part of the dataset. We thank Kamryn Almas for her proof-reading.

This work was made possible by the funding of the GPU by the Domaine d'Intérêt Majeur "Science du Texte et Connaissances Nouvelles" as well as the funding for dataset production and evaluation of the Domaine d'Intérêt Majeur "Matériaux Anciens et Patrimoine" in the context of the *CREMMA* project. Finally, we thank our colleagues from INRIA for the availability of the *eScriptorium* test interface and specifically Ben Kiessling for his availability for troubleshooting.

References

Michele Alberti, Manuel Bouillon, Rolf Ingold, and Marcus Liwicki. Open Evaluation Tool for Layout Analysis of Document Images. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pages 43–47, Kyoto, Japan, nov 2017. ISBN 978-1-5386-3586-5. doi: 10.1109/ICDAR.2017.311.

Jean-Pierre Aniel. MANDRAGORE. Une base de données iconographiques sur les manuscrits de la Bibliothèque nationale de Paris. *Le médiéviste et l'ordinateur*, 26(1):18–20, 1992. doi: 10.3406/medio.1992.1369. URL https://www.persee.fr/doc/medio_0223-3843_1992_num_26_1_1369. Publisher: Persée - Portail des revues scientifiques en SHS.

Sanae Boutarfass and Bernard Besserer. Improving CNN-based colorization of black and white photographs. In *2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS)*, pages 96–101, December 2020. doi: 10.1109/IPAS50080.2020.9334930.

Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 2020. ISSN 2078-2489. doi: 10.3390/info11020125. URL https://www.mdpi.com/2078-2489/11/2/125.

Jean-Baptiste Camps. *La Chanson d'Otinel : édition complète du corpus manuscrit et prolégomènes à l'édition critique – digital appendices*. PhD thesis, Université Paris IV, December 2017. URL https://zenodo.org/record/1116736#.XN1ufC3M00Q.

Jean-Baptiste Camps, Thibault Clérice, and Ariane Pinche. Stylometry for Noisy Medieval Data: Evaluating Paul Meyer's Hagiographic Hypothesis. *arXiv:2012.03845 [cs]*, December 2020. URL http://arxiv.org/abs/2012.03845. arXiv: 2012.03845.

Jean-Baptiste Camps, Chahan Vidal-Gorène, and Marguerite Vernet. Handling Heavily Abbreviated Manuscripts: HTR engines vs text normalisation approaches, May 2021. URL https://hal-enc.archives-ouvertes.fr/hal-03279602.

Yu Chen, Yeyun Luo, Youdong Ding, and Bing Yu. Automatic Colorization of Images from Chinese Black and White Films Based on CNN. In *2018 International Conference on Audio, Language and Image Processing (ICALIP)*, pages 97–102, July 2018. doi: 10.1109/ICALIP.2018.8455654.

Christian Clausner, Stefan Pletschacher, and Apostolos Antonacopoulos. Scenario driven in-depth performance evaluation of document layout analysis methods. In *2011 International Conference on Document Analysis and Recognition*, pages 1404–1408. IEEE, 2011.

Thibault Clérice. Evaluating Deep Learning Methods for Word Segmentation of Scripta Continua Texts in Old French and Latin. working paper or preprint, June 2019. URL https://hal.archives-ouvertes.fr/hal-02154122.

Thibault Clérice and Ariane Pinche. Choco-Mufin, a tool for controlling characters used in OCR and HTR projects, 9 2021a. URL https://github.com/PonteIneptique/choco-mufin.

Thibault Clérice and Ariane Pinche. HTRVX, HTR Validation with XSD, 9 2021b. URL https://github.com/HTR-United/HTRVX.
Council on Library & Information Resources. APPENDIX VI: Comparative Costs for Book Treatments. In *The Evidence in Hand: Report of the Task Force on the Artifact in Library Collections*, volume 109. Council on Library & Information Resources, November 2001. URL https://www.clir.org/pubs/reports/pub103/appendix6/.
Samuel Goree. The Limits of Colorization of Historical Images by AI, April 2021. URL https://hyperallergic.com/639395/the-limits-of-colorization-of-historical-images-by-ai/.
Louis Holtz. Les premières années de l'Institut de recherche et d'histoire des textes. *La revue pour l'histoire du CNRS*, 2000(2), May 2000. ISSN 1298-9800. doi: 10.4000/histoire-cnrs.2742. URL http://journals.openedition.org/histoire-cnrs/2742.
Madhab Raj Joshi, Lewis Nkenyereye, Gyanendra Prasad Joshi, S. M. Riazul Islam, Mohammad Abdullah-Al-Wadud, and Surendra Shrestha. Auto-Colorization of Historical Images Using Deep Convolutional Neural Networks. *Mathematics*, 8(12):2258, December 2020. doi: 10.3390/math8122258. URL https://www.mdpi.com/2227-7390/8/12/2258. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.
Gwen C. Katz. Colorization APIs are becoming widespread..., April 2021. URL https://twitter.com/gwenckatz/status/1381652071695351810.
Benjamin Kiessling. Kraken - an Universal Text Recognizer for the Humanities. UItrecht, July 2019. CLARIAH. URL https://dev.clariah.nl/files/dh2019/boa/0673.html.
Benjamin Kiessling. A Modular Region and Text Line Layout Analysis System. pages 313–318. IEEE Computer Society, September 2020. ISBN 978-1-72819-966-5. doi: 10.1109/ICFHR2020.2020.00064. URL https://www.computer.org/csdl/proceedings-article/icfhr/2020/996600a313/1p2VtgcXVLi.
Benjamin Kiessling, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. eScriptorium: An Open Source Platform for Historical Document Analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 19–19, September 2019. doi: 10.1109/ICDARW.2019.10032.
L. W. C. van Lit. *The Digital Materiality of Digitized Manuscripts*. Brill, October 2019. ISBN 978-90-04-40035-1. doi: 10.1163/9789004400351_004. URL http://brill.com/view/book/9789004400351/BP000011.xml. Pages: 51-72 Publication Title: Among Digitized Manuscripts. Philology, Codicology, Paleography in a Digital World Section: Among Digitized Manuscripts. Philology, Codicology, Paleography in a Digital World.
Enrique Manjavacas, Mike Kestemont, and F.B. Karsdorp. *A Robot's Street Credibility: Modeling authenticity judgments for artificially generated Hip-Hop lyrics*. 2019.
Allardyce Nicoll. The "Basic" Use of Microfilm. *PMLA*, 68(4):62–66, 1953. ISSN 0030-8129. URL http://www.jstor.org/stable/2698986. Publisher: Modern Language Association.
Ariane Pinche. CREMMA Medieval, an Old French dataset for HTR and segmentation, 8 2021a. URL https://github.com/htR-United/cremma-medieval.
Ariane Pinche. *Edition nativement numérique du recueil hagiographique "Li Seint Confessor" de Wauchier de Denain d'après le manuscrit 412 de la Bibliothèque nationale de France*. Thèse de doctorat, Lyon, Lyon, May 2021b. URL http://www.theses.fr/s150996.
British Manuscripts Project. British Manuscripts Project. *PMLA*, 59:1463–1488, 1944. ISSN 0030-8129. URL https://www.jstor.org/stable/459214. Publisher: Modern Language Association.
S. K. Stevens. The Use of Microfilm. *Oregon Historical Quarterly*, 51(3):167–179, 1950. ISSN 0030-4727. URL https://www.jstor.org/stable/20611982. Publisher: Oregon Historical Society.
Dominique Stuzmann. Paléographie statistique pour décrire, identifier, dater. . . normaliser pour coopérer et aller plus loin ? In Franz Fischer, Christiane Fritze, and Georg Vogeler, editors, *Kodikologie und Paläographie im digitalen Zeitalter 2 - Codicology and Palaeography in the Digital Age 2*, volume 3, pages 247–277. Books on Demand (BoD), Norderstedt, 2011. URL https://kups.ub.uni-koeln.de/4353/.
Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. Instance-aware image colorization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

A DATASET SAMPLES

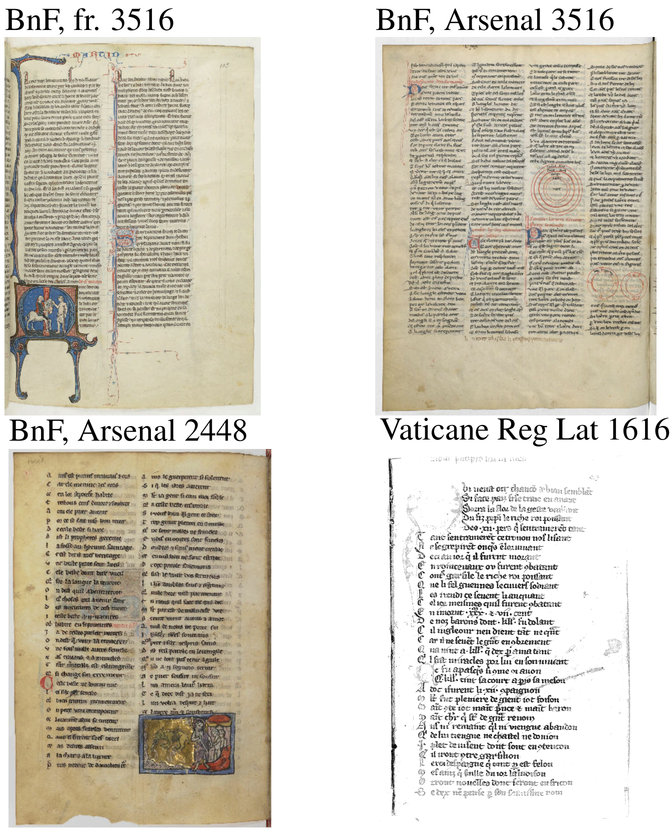

Figure 13: Samples from the HTR and segmentation training dataset

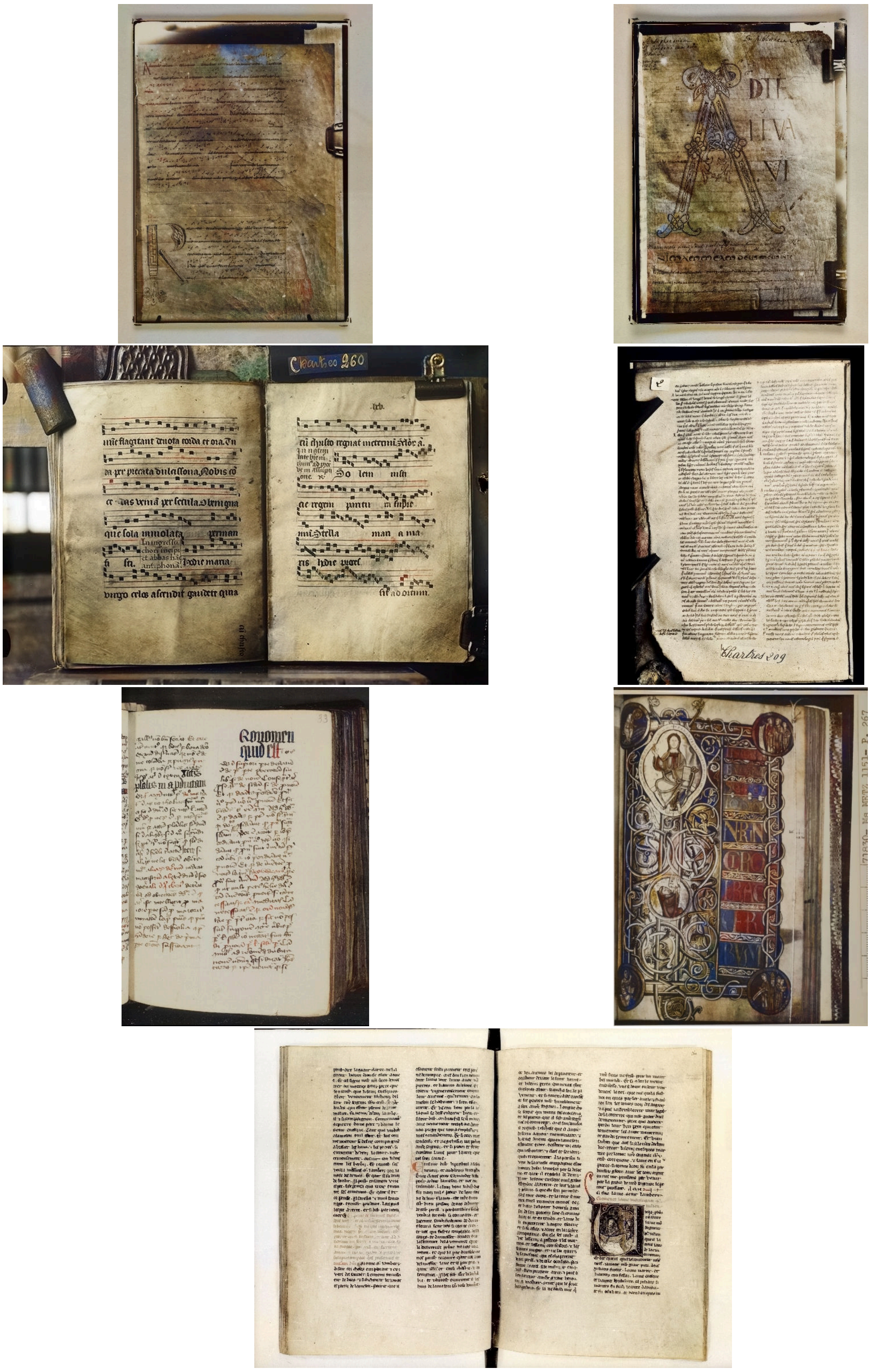

Figure 14: Samples from colorization output on microfilms. First 4 are manuscripts from Chartes (0047, 0260, 0209), the following two are from Metz (0643 and 1151), the last if from our testing set (BnF fr. 17229)