

# DocEnTr: An End-to-End Document Image Enhancement Transformer

Mohamed Ali Souibgui<sup>§</sup>

Computer Vision Center

Universitat Autònoma de Barcelona

Barcelona, Spain

[msouibgui@cvc.uab.es](mailto:msouibgui@cvc.uab.es)

Sanket Biswas<sup>§</sup>

Computer Vision Center

Universitat Autònoma de Barcelona

Barcelona, Spain

[sbiswas@cvc.uab.es](mailto:sbiswas@cvc.uab.es)

Sana Khamekhem Jemni<sup>§</sup>

Digital Research Center of Sfax

MIRACL Laboratory, University of Sfax

Sfax, Tunisia

[sana.khamekhem@gmail.com](mailto:sana.khamekhem@gmail.com)

Yousri Kessentini

Digital Research Center of Sfax

SM@RTS Laboratory

Sfax, Tunisia

[yousri.kessentini@crns.rnrt.tn](mailto:yousri.kessentini@crns.rnrt.tn)

Alicia Fornés, Josep Lladós

Computer Vision Center, Computer Science Dept.

Universitat Autònoma de Barcelona

Barcelona, Spain

[{afornes, josep}@cvc.uab.es](mailto:{afornes, josep}@cvc.uab.es)

Umapada Pal

CVPR Unit

Indian Statistical Institute

Kolkata, India

[umapada@isical.ac.in](mailto:umapada@isical.ac.in)

**Abstract**—Document images can be affected by many degradation scenarios, which cause recognition and processing difficulties. In this age of digitization, it is important to denoise them for proper usage. To address this challenge, we present a new encoder-decoder architecture based on vision transformers to enhance both machine-printed and handwritten document images, in an end-to-end fashion. The encoder operates directly on the pixel patches with their positional information without the use of any convolutional layers, while the decoder reconstructs a clean image from the encoded patches. Conducted experiments show a superiority of the proposed model compared to the state-of-the-art methods on several DIBCO benchmarks. Code and models will be publicly available at: <https://github.com/dali92002/DocEnTR>.

## I. INTRODUCTION

The preservation and legibility of document images (especially the historical ones) are of utmost priority for the Document Image Analysis and Recognition (DIAR) research. Document records usually contain significant information and in the historical cases it dates back centuries and decades [1]. The conservation of document records can be hampered by several kinds of degradation such as smears, stains, artefacts, pen strokes, bleed-through effects and uneven illumination. These distortions could heavily impact the subsequent downstream tasks for information processing, such as segmentation, Optical Character Recognition (OCR), information spotting and layout analysis. This manifests the need for a robust pre-processing task that denoises and reconstructs a high-quality clean image from its already degraded counterpart. Document Image Enhancement (DIE) aims towards restoring the quality of the degraded document samples to yield a clear enhanced version that is locally uniform.

In recent times, Convolutional Neural Network (CNN)-based approaches have been widely applied to DIE related sub-tasks, like binarization [2], [3], deblurring [4], shadow [5] and

watermark removal [6], etc. Although the performance of these models has significantly improved over classical handcrafted techniques, they do have their own set of drawbacks. Firstly, CNNs operate on regular grids and using the same convolutional filter to restore different regions of a degraded document image may not be a sensible choice. Secondly, CNNs fail to capture high-level long-range dependencies as they are more suited for extracting low-level spatial information from images.

With the recent success of transformers in Natural Language Processing (NLP) [7], [8], its application to computer vision problems (like image recognition [9], object detection [10], visual question answering [11], handwritten text recognition (HTR) [12], etc.) also started getting more prominence. The self-attention mechanism proposed in [7] helps to capture global interactions between contextual features. Using local information combined with the knowledge of long-range global spatial arrangement is beneficial for an efficient image restoration model. This local information is often encoded in the patch content of an image and the large scale organization is contained in the redundancy of this information across the patches of the image [13]. Contrary to CNNs, which process pixel arrays, Vision Transformers (ViTs) [9] split an image into fixed-size patches (eg. 8x8, 16x16 etc.), they correctly embeds each of them as latent representation, and include positional embedding information as input to the transformer encoder. This allows to encode the relative location of the patches, along with both local (spatial) and global (semantic) long-range dependencies. The motivation of using ViTs for our overall proposed baseline model is that a missing/degraded patch in the distorted document image can be recovered from the neighbouring patches information with the power of the multi-head self-attention in ViTs, which quantifies pairwise global reasoning between them. Also, ViTs have been adapted in the overall model pipeline in an encoder-decoder based setting, inspired by the concept of denoising autoencoders

<sup>§</sup>Equal contribution

[14] used in reconstruction of corrupted input data. The encoder is mapping the degraded image patches into latent representations, whereas the decoder is recovering a clean image version from those encoded representations.

The overall contributions of our work can be summarized into three folds:

- We introduce a simple and flexible Document image Enhancement Transformer (DocEnTr), an end-to-end image enhancement approach, that effectively restores and enhances a degraded document image provided as input. As far as we know, DocEnTr is the first pure transformer-based baseline that leverages the effectiveness of Vision Transformers (ViTs) in an encoder-decoder based framework, without any dependency on CNNs.
- We have addressed document binarization as the key problem study in this work to investigate the power of DocEnTr architecture. Experimental evaluation shows that DocEnTr achieves state-of-the-art results on standard document binarization benchmarks (DIBCO), for both machine-printed and handwritten degraded document images.
- A comprehensive and intuitive case study has been dedicated in Section IV to prove the utility of ViTs with its multi-headed self-attention mechanism in the task of document enhancement.

The rest of this paper is organized as follows. In Section II we review the state of the art. The Document image Enhancement Transformer (DocEnTr) is described in Section III. Section IV contains an analysis of the extensive experimentation that has been conducted, including different quantitative and qualitative studies. Finally, in Section V we draw the conclusions and propose open challenges for future research directions.

## II. RELATED WORK

### A. Document Image Enhancement

This work is an application within the DIE, which has been an active field within the DIAR community. The first classic methods were based on thresholding, which means finding a single (global) or multiple (local) threshold(s) value(s) for the document. These threshold values are used to classify the document image pixels into foreground (black) or background (white) [15], [16]. These methods are still evolving in the recent years using machine learning tools, for instance, with support vector machines (SVM) [17]. Later, energy based methods were introduced. These are based on tracking the text pixels by maximizing its energy function [18], while minimizing the one of the degraded background. However, the results using those approaches were unsatisfactory [19].

Recently, deep learning based methods were used to tackle this problem by learning the enhancement directly from raw data. In [20], the problem was formulated as pixels classification. Each pixel is classified as black or white depending on a sequence of the surrounding pixels, where a 2D Long Short-Term Memory (LSTM) was trained for this task. This process

is, of course, time consuming. A more practical solution is to map the images from the degraded domain to the enhanced one in an end-to-end fashion with CNN auto-encoders. These latter, hence, were leading the recent improvements in image denoising [21] and more particularly documents enhancement tasks, like binarization [22], [23], [24], deblurring problems [4] and so on. Following this strategy, a fully CNN model was proposed in [25] to binarize the degraded document images at multiple image scales. Similarly, [2] proposed an auto-encoder architecture that performs a cascade of pre-trained U-Net models [26] to learn the binarization using less amount of data. Moreover, generation models (GAN) were employed for this task to generate clean images by conditioning on the degraded versions. These architectures are composed of a generative model that produces a clean version of the image and a discriminator to assess the binarization result. Both models are usually composed of fully (or partially) CNN layers. In [6], a conditional GAN approach was proposed for different enhancement tasks achieving good results in document images cleaning, binarization, deblurring and dense watermarks removal. This method was recently extended in [3] by adding a second discriminator to assess the text readability for the goal of obtaining an enhanced image that is clean and readable at the same time. A similar cGAN's based method was also proposed in [27], [28], [29], [30].

### B. Transformers in Vision and Image Enhancement Tasks

In the very recent years, transformers are behind the advances in deep learning applications. Transformer based architectures firstly showed a great success in NLP tasks [7], [8] for text translation and embedding, surpassing the previous LSTM approaches. This motivates many works to employ them for the vision tasks, for instance, classification [9], object detection [10], document understanding [31], [32], [33], etc. More related to this paper, transformers were also used for natural image restoration [34] and document images dewarping [35]. However, the architectures that were used in these later image and document enhancement approaches are still relying on the CNN feature extractors before passing to the transformers stage. Also, the CNN are used to reconstruct the output image. Contrary, what we are proposing in this work is a fully transformer approach that attends directly to the patches on the input images and reconstruct the pixels without the using of any CNN layer.

## III. METHOD

The proposed model is a scalable auto-encoder that uses vision transformers in its encoder and decoder parts, as illustrated in Fig 1. The degraded image is first divided into patches before entering to the encoder part. During encoding, the patches are mapped to a latent representation of tokens, where each token is associated with a degraded patch. Then, the tokens are passed to the decoder that outputs the enhanced version of patches. Unlike the CNN based auto-encoders, which were usually employed for the document image enhancement tasks, the transformer auto-encoder is

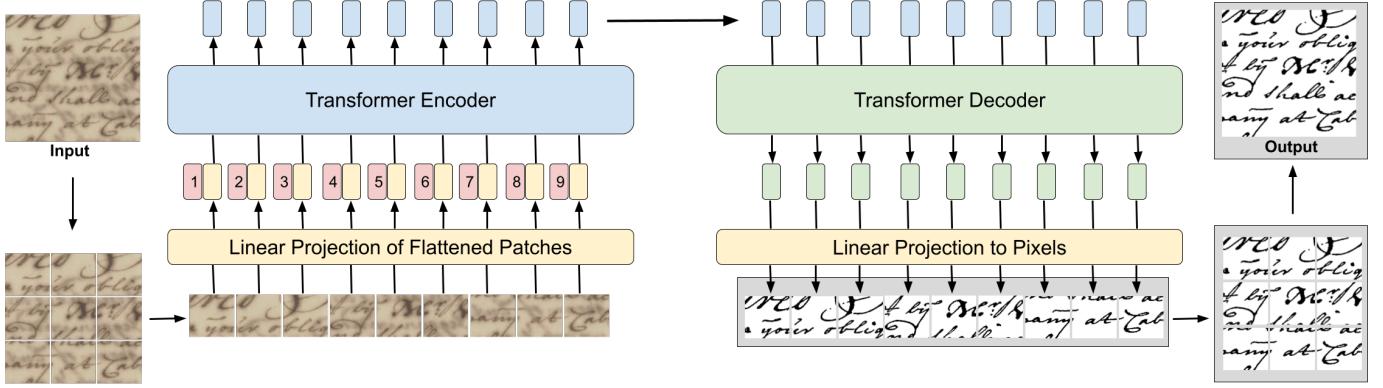


Fig. 1. Proposed model: The input image is split into patches, which are linearly embedded, and the position information are added to them. The resulting sequence of vectors are fed to a standard Transformer encoder to obtain the latent representations. These representations are fed to another Transformer representing the decoder to obtain the decoded vector, which is linearly projected to vectors of pixels representing the output image patches.

profiting from the self attention mechanism which gives a global information during every patch enhancement. Both decoder and especially encoder are inspired from the vision transformer (ViT) [9] architecture. We present more details of the model's architecture in what follows.

#### A. Encoder

In the encoding stage (left part of Fig.1), given an image, we divide it into a set of patches. Then, we embed these patches to obtain the tokens and add their positional information. After that, a number of transformer blocks is employed to map these tokens into the encoded latent representation. These blocks follow the same structure as [9], composed of alternating layers of multi-headed self-attention and multi-layered perceptron (MLP). Each of these blocks are preceded by a LayerNorm (LN) [36], and followed by a residual connection. The patches embedding size and the number of transformer blocks are set depending on the model size.

#### B. Decoder

The decoder part consists of a series of transformer blocks (having the same number as the encoder blocks) that take as an input the sequence of outputted tokens from the encoder. These tokens are propagated in the transformer decoder blocks, and then projected with a linear layer to the desired pixel values. This makes each element of the output correspond to a vector representing a flattened patch in the output image. The ground truth pixel values are obtained by dividing the ground truth (GT) clean image into patches (in the same way as the input degraded image) and flattening them into vectors. A mean squared error (MSE) loss is used between the model's output and the GT pixel patches to train the model.

#### C. Model Variants

Following a similar convention as previous works [8], [9], the proposed model configuration can be modified to produce different variants. In our experiments we define three types of variants which are "Small", "Base" and "Large", as enlisted in Table I. Evidently, setting a larger model require more

computational memory and training time since the number of model parameters is increasing. Thus, a trade off between the model size and its enhancement performance must be taken into consideration.

TABLE I  
DETAILS OF OUR MODEL VARIANTS

Model	Layers	Dim	Attention Heads	# Parameters
DocEnTr-Small	6	512	4	17M
DocEnTr-Base	12	768	8	68M
DocEnTr-Large	24	1024	16	255M

#### IV. EXPERIMENTAL VALIDATION

To validate our model, we use the datasets proposed in the different DIBCO and H-DIBCO contests [37] for printed and handwritten degraded document images binarization and compare our results with the state of the art methods. Before these experiments, we have conducted different investigations for a proper selection of the hyperparameters.

#### A. Choosing the Best Model Configuration

We begin our experiments by choosing the configuration that gives the best performance from our model variants (Small, Base or Large). For training, each degraded image and its GT clean one is divided into overlapped patches with sizes  $256 \times 256 \times 3$ , the overlapping was set vertically and horizontally by a half of the patches size (means 128). These resultant images (patches) will be used by our models as an input and expected output (training data). For results evaluation, and same as the usual approaches [38], we utilize the following metrics: Peak signal-to-noise ratio (PSNR), F-Measure (FM), pseudo-F-measure ( $F_{ps}$ ) and Distance reciprocal distortion metric (DRD). We used in this experiment the DIBCO 2017 dataset, and the obtained results are given in Table II. As it can be seen, a larger model gives a better result in all the metrics, but it requires more computation resources. Thus, we recommend using a Base model for a binarization

task. Nevertheless, we will test as well the Large version in following experiments.

TABLE II  
RESULTS OF VARYING THE MODEL SIZE FOR THE DIBCO 2017 DATASET.  
↑: THE HIGHER THE BETTER. ↓: THE LOWER THE BETTER.

Model	PSNR ↑	FM ↑	$F_{ps}$ ↑	DRD ↓
DocEnTr-Small	18.29	91.06	93.82	2.78
DocEnTr-Base	18.69	91.66	94.11	2.63
DocEnTr-Large	<b>18.85</b>	<b>92.14</b>	<b>94.58</b>	<b>2.53</b>

Next, we do another experiment related to the input image size, and the patches size that are used by our model. The reason behind is that having different image size and patch size can affect the binarization since the model is accessing to different type of information (from global to local). The obtained results using the Base model are given in Table III. As it can be seen, a slightly better performance is obtained using an input with the smaller size ( $256 \times 256 \times 3$  compared to  $512 \times 512 \times 3$ ). However, we can notice that the performance is highly improved when using a smaller patch size. The reason is that, by employing a smaller patch size, we make each patch of the image attending to more and much local patches during the self-attention. Thus, the model is looking to more and much fine information during the enhancement process with  $8 \times 8$  patch size. But, as before, using a smaller patch size means augmenting the model parameters, requiring more computation resources.

TABLE III  
RESULTS OF VARYING THE INPUT AND PATCH SIZES FOR THE DIBCO 2017 DATASET

Input Size	Patch Size	PSNR ↑	FM ↑	$F_{ps}$ ↑	DRD ↓
$256 \times 256 \times 3$	$8 \times 8$	<b>19.11</b>	<b>92.53</b>	<b>95.15</b>	<b>2.37</b>
$256 \times 256 \times 3$	$16 \times 16$	18.69	91.66	94.11	2.63
$256 \times 256 \times 3$	$32 \times 32$	17.57	89.37	91.99	3.44
$512 \times 512 \times 3$	$8 \times 8$	18.91	92.2	94.93	2.45
$512 \times 512 \times 3$	$16 \times 16$	18.66	92.15	93.89	2.54
$512 \times 512 \times 3$	$32 \times 32$	17.27	89.43	91.51	3.54

### B. Quantitative Evaluation

After choosing the best hyper-parameters of the model, we conduct the experiments on the different datasets and compare our result with the related approaches. We begin by testing with the DIBCO 2011 dataset [39]. This dataset contains degraded document images with handwritten and printed text. For training, we use all the images from the other DIBCO and H-DIBCO datasets (except DIBCO 2019) and the Palm Leaf dataset [40]. These images are split into overlapped images with size  $256 \times 256 \times 3$  before being fed to the model. The obtained results are given in Table IV, where we can notice a superiority of our method compared to the different variations of the related approaches. We choose to compare with different families of approaches: classic thresholding and deep learning based methods (whether basing on CNN or cGAN). Our model

DocEnTr-Base{8}, which means using the Base setting with a patch size of  $8 \times 8$ , gives the best PSNR and DRD compared to all the other methods. While the model DocEnTr-Large{16}, which means using the Large setting with a patch size of  $16 \times 16$ , leads to the second best performance in the metrics PSNR,  $F_{ps}$  and DRD. We note that for a computation reason, we were not able to train the Large setting with a patch size of  $8 \times 8$ .

TABLE IV  
COMPARATIVE RESULTS OF OUR PROPOSED METHOD ON DIBCO 2011 DATASET. THRESH: THRESHOLDING, TR: TRANSFORMERS.

Method	Model	PSNR ↑	FM ↑	$F_{ps}$ ↑	DRD ↓
Otsu [15]	Thres.	15.70	82.10	—	9.00
Savoula et al. [16]	Thres.	15.60	82.10	—	8.50
Vo et al. [41]	CNN	20.10	93.30	—	2.00
Kang et al [2]	CNN	19.90	<b>95.50</b>	—	1.80
Tensmeyer et al [25]	CNN	20.11	93.60	<b>97.70</b>	1.85
Zhao et al. [41]	cGAN	20.30	93.80	—	1.80
<b>DocEnTr-Base{8}</b>	Tr	<b>20.81</b>	94.37	96.15	<b>1.63</b>
<b>DocEnTr-Base{16}</b>	Tr	20.11	93.48	96.12	1.93
<b>DocEnTr-Large{16}</b>	Tr	20.62	94.24	96.71	1.69

After that, we test our model on the H-DIBCO 2012 dataset [42], which contains degraded handwritten document images. As in the previous experiment, we use the other datasets for training with the same split size. The obtained results are shown in Table V, where we can notice that our model gives the best performance in terms of PSNR and FM with the Base{8} configuration. We notice also that the other configuration gives competitive results compared to the other approaches.

TABLE V  
COMPARATIVE RESULTS OF OUR PROPOSED METHOD ON H-DIBCO 2012 DATASET. THRESH: THRESHOLDING, TR: TRANSFORMERS.

Method	Model	PSNR ↑	FM ↑	$F_{ps}$ ↑	DRD ↓
Otsu [15]	Thres.	15.03	80.18	82.65	26.46
Savoula et al. [16]	Thres.	16.71	82.89	87.95	6.59
Kang et al [2]	CNN	21.37	95.16	96.44	<b>1.13</b>
Tensmeyer et al [25]	CNN	20.60	92.53	<b>96.67</b>	2.48
Zhao et al. [41]	cGAN	21.91	94.96	96.15	1.55
Jemni et al. [3]	cGAN	22.00	95.18	94.63	1.62
<b>DocEnTr-Base{8}</b>	Tr	<b>22.29</b>	<b>95.31</b>	96.29	1.60
<b>DocEnTr-Base{16}</b>	Tr	21.03	93.31	94.72	2.31
<b>DocEnTr-Large{16}</b>	Tr	22.04	95.09	96.00	1.64

Moreover, we tested with the more recent DIBCO 2017 dataset. In this dataset our model achieves the best performance in all the evaluation metrics, as presented in Table VI.

Lastly, we test on the H-DIBCO 2018 dataset. Here, as shown in Table VII, the best performance is achieved by [3] basing on cGAN. Anyway, we can notice that our model is still very competitive since it ranks second in the PSNR, FM and  $F_{ps}$  metrics.

To summarize the quantitative evaluation, we demonstrate that our model gives good results compared to the state of the art approaches. This was shown by obtaining the best results in most of the evaluation metrics with the H-DIBCO 2011, DIBCO 2012 and DIBCO 2017 benchmarks.

TABLE VI  
COMPARATIVE RESULTS OF OUR PROPOSED METHOD ON DIBCO 2017  
DATASET. THRESH: THRESHOLDING, TR: TRANSFORMERS.

Method	Model	PSNR $\uparrow$	FM $\uparrow$	$F_{ps}$ $\uparrow$	DRD $\downarrow$
Otsu [15]	Thres.	13.85	77.73	77.89	15.54
Savoula et al. [16]	Thres.	14.25	77.11	84.1	8.85
Kang et al [2]	CNN	15.85	91.57	93.55	2.92
Competition top [19]	CNN	18.28	91.04	92.86	3.40
Zhao et al. [41]	cGAN	17.83	90.73	92.58	3.58
Jemni et al. [3]	cGAN	17.45	89.8	89.95	4.03
<b>DocEnTr-Base{8}</b>	Tr	<b>19.11</b>	<b>92.53</b>	<b>95.15</b>	<b>2.37</b>
<b>DocEnTr-Base{16}</b>	Tr	18.69	91.66	94.11	2.63
<b>DocEnTr-Large{16}</b>	Tr	18.85	92.14	94.58	2.53

TABLE VII  
COMPARATIVE RESULTS OF OUR PROPOSED METHOD ON DIBCO 2018  
DATASET. THRESH: THRESHOLDING, TR: TRANSFORMERS.

Method	Model	PSNR $\uparrow$	FM $\uparrow$	$F_{ps}$ $\uparrow$	DRD $\downarrow$
Otsu [15]	Thres.	9.74	51.45	53.05	59.07
Savoula et al. [16]	Thres.	13.78	67.81	74.08	17.69
Kang et al [2]	CNN	19.39	89.71	91.62	<b>2.51</b>
Competition top [19]	CNN	19.11	88.34	90.24	4.92
Zhao et al. [41]	cGAN	18.37	87.73	90.60	4.58
Jemni et al. [3]	cGAN	<b>20.18</b>	<b>92.41</b>	<b>94.35</b>	2.60
<b>DocEnTr-Base{8}</b>	Tr	19.46	90.59	93.97	3.35
<b>DocEnTr-Base{16}</b>	Tr	19.33	89.97	93.5	3.68
<b>DocEnTr-Large{16}</b>	Tr	19.47	89.21	92.54	3.96

### C. Qualitative Evaluation

After presenting the achieved quantitative results by our model, we present in this subsection some qualitative results. We begin by showing the enhancing performance of our method. This is illustrated in Fig. 2, where we compare our binarization results with the GT clean images. As it can be seen, our model produces highly clean images, which are very close to the optimal GT images, reflecting the good quantitative performance that was obtained in the previous subsection.

Then, we present a quantitative comparison of our method with the related approaches. This is shown in Fig. 3, where we can notice the superiority of our model in recovering a highly degraded image over the classic thresholding [15], [16], CNN [2], and cGAN [3] methods.

### D. Self-attention Mechanism

As we stated above, our method differs from the CNN related ones by employing the transformers to enhance the degraded document images. The self-attention mechanism used in the transformer blocks gives a global view to every token on the other tokens that represents the patches within the image for a better enhancing result. A visual illustration of the attention maps of the last layer from the encoder is given in Fig. 4. As it can be seen, a token can attend to all the patches within the image. In these test cases each token (patch representation) is focusing on the text elements, while ignoring the degraded patches. Thus, the attending patches are decoded later and projected to pixels while taking into consideration a high-level global information from the attended neighbouring patches that cover the full input image. We also notice that



Fig. 2. Qualitative results of our proposed method in binarization of some samples from the DIBCO and H-DIBCO datasets. Images in columns are: Left: original image, Middle: GT image, Right: Binarized image using our proposed method.

the attention maps are mostly matching the text of the GT images, which lead to a satisfactory binarization result that is closer to the GT. This supports the utility of using the transformers with its powerful self-attention mechanism in the image enhancement task. However, in other sample cases as illustrated in Fig. 5, we observe that the attention maps are considering some portions of the text as a background

Inhalt.	
1. Vom Gesicht.	Seite 342
2. Vom Grunde.	344
3. Erklärungen seltsamer Gefäße; Anwendungen	
4. wunderlicher Begegnen der Schwangeren und	
5. Hysterischen.	350
6. Von den Leidenschaften, und der Notwendigkeit	
7. des Studiums der Menschenentnthalzung ic.	364
8. Von physischen Geheimnissen oder den Wissen:	
9. Höfsten der Sphären.	368
10. Von sonderlichen Gefühlen und Empfindungen.	375
11. Theorie angenehmer Empfindungen.	378

Original

Inhalt.	
1. Vom Gesicht.	Seite 342
2. Vom Grunde.	344
3. Erklärungen seltsamer Gefäße; Anwendungen	
4. wunderlicher Begegnen der Schwangeren und	
5. Hysterischen.	350
6. Von den Leidenschaften, und der Notwendigkeit	
7. des Studiums der Menschenentnthalzung ic.	364
8. Von physischen Geheimnissen oder den Wissen:	
9. Höfsten der Sphären.	368
10. Von sonderlichen Gefühlen und Empfindungen.	375
11. Theorie angenehmer Empfindungen.	378

Otsu [15]

Inhalt.	
1. Vom Gesicht.	Seite 342
2. Vom Grunde.	344
3. Erklärungen seltsamer Gefäße; Anwendungen	
4. wunderlicher Begegnen der Schwangeren und	
5. Hysterischen.	350
6. Von den Leidenschaften, und der Notwendigkeit	
7. des Studiums der Menschenentnthalzung ic.	364
8. Von physischen Geheimnissen oder den Wissen:	
9. Höfsten der Sphären.	368
10. Von sonderlichen Gefühlen und Empfindungen.	375
11. Theorie angenehmer Empfindungen.	378

Jemni et al. [3]

Inhalt.	
1. Vom Gesicht.	Seite 342
2. Vom Grunde.	344
3. Erklärungen seltsamer Gefäße; Anwendungen	
4. wunderlicher Begegnen der Schwangeren und	
5. Hysterischen.	350
6. Von den Leidenschaften, und der Notwendigkeit	
7. des Studiums der Menschenentnthalzung ic.	364
8. Von physischen Geheimnissen oder den Wissen:	
9. Höfsten der Sphären.	368
10. Von sonderlichen Gefühlen und Empfindungen.	375
11. Theorie angenehmer Empfindungen.	378

Competition Winner

Fig. 3. Qualitative results of the different binarization methods on the sample number 12 from DIBCO 2017 Dataset.

region. Hence, the resultant enhanced image is removing foreground text because it considers it as a background noise. This explains the failure of the self-attention paradigm in these scenarios.

## V. CONCLUSION

This paper presents a novel transformer-based architecture called DocEnTr for document image enhancement. To the best of our knowledge, this is the first pure transformer model addressing DIE related problems. The model captures high-level global long-range dependencies using the self-attention mechanism for a better performance. Quantitative and qualitative results on the DIBCO benchmarks prove the effectiveness of DocEnTr in recovering highly degraded document images. It is a simple and flexible framework that can also be easily applied to enhance other kinds of degradation occurring in document images (like blur, shadow, warps, stains etc). These aspects will be investigated in a future work. We also wish to investigate a self-supervised learning stage that can substantially benefit from large amounts of unlabeled data.

Inhalt.	
1. Vom Gesicht.	Seite 342
2. Vom Grunde.	344
3. Erklärungen seltsamer Gefäße; Anwendungen	
4. wunderlicher Begegnen der Schwangeren und	
5. Hysterischen.	350
6. Von den Leidenschaften, und der Notwendigkeit	
7. des Studiums der Menschenentnthalzung ic.	364
8. Von physischen Geheimnissen oder den Wissen:	
9. Höfsten der Sphären.	368
10. Von sonderlichen Gefühlen und Empfindungen.	375
11. Theorie angenehmer Empfindungen.	378

Ground Truth

Inhalt.	
1. Vom Gesicht.	Seite 342
2. Vom Grunde.	344
3. Erklärungen seltsamer Gefäße; Anwendungen	
4. wunderlicher Begegnen der Schwangeren und	
5. Hysterischen.	350
6. Von den Leidenschaften, und der Notwendigkeit	
7. des Studiums der Menschenentnthalzung ic.	364
8. Von physischen Geheimnissen oder den Wissen:	
9. Höfsten der Sphären.	368
10. Von sonderlichen Gefühlen und Empfindungen.	375
11. Theorie angenehmer Empfindungen.	378

Sauvola et al. [16]

Inhalt.	
1. Vom Gesicht.	Seite 342
2. Vom Grunde.	344
3. Erklärungen seltsamer Gefäße; Anwendungen	
4. wunderlicher Begegnen der Schwangeren und	
5. Hysterischen.	350
6. Von den Leidenschaften, und der Notwendigkeit	
7. des Studiums der Menschenentnthalzung ic.	364
8. Von physischen Geheimnissen oder den Wissen:	
9. Höfsten der Sphären.	368
10. Von sonderlichen Gefühlen und Empfindungen.	375
11. Theorie angenehmer Empfindungen.	378

Kang et al. [2]

Inhalt.	
1. Vom Gesicht.	Seite 342
2. Vom Grunde.	344
3. Erklärungen seltsamer Gefäße; Anwendungen	
4. wunderlicher Begegnen der Schwangeren und	
5. Hysterischen.	350
6. Von den Leidenschaften, und der Notwendigkeit	
7. des Studiums der Menschenentnthalzung ic.	364
8. Von physischen Geheimnissen oder den Wissen:	
9. Höfsten der Sphären.	368
10. Von sonderlichen Gefühlen und Empfindungen.	375
11. Theorie angenehmer Empfindungen.	378

Ours

Degraded	Attention	Predicted	GT

Fig. 4. Attention maps from the 2<sup>nd</sup> head of the last layer of DocEnTr{8} encoder. We display the self-attention for different (random) tokens.

Degraded	Attention	Predicted	GT

Fig. 5. Attention maps from the 2<sup>nd</sup> head of the last layer of DocEnTr{8} encoder. We display the self-attention for different (random) tokens. (A failure case).

## ACKNOWLEDGMENT

This work has been partially supported by the Swedish Research Council (grant 2018-06074, DECRYPT), the Spanish projects RTI2018-095645-B-C21, the CERCA Program / Generalitat de Catalunya, the FCT-19-15244, the Catalan projects 2017-SGR-1783, PhD Scholarship from AGAUR (2021FIB-10010) and DocPRESERV project (Swedish STINT grant).

## REFERENCES

- [1] B. Megyesi, N. Blomqvist, and E. Pettersson, “The decode database: Collection of historical ciphers and keys,” in *The 2nd International Conference on Historical Cryptology, HistoCrypt 2019, June 23-26 2019, Mons, Belgium*, 2019, pp. 69–78.
- [2] S. Kang, B. K. Iwana, and S. Uchida, “Complex image processing with less data document image binarization by integrating multiple pretrained u-net modules,” *Pattern Recognition*, vol. 109, p. 107577, 2021.
- [3] S. K. Jemni, M. A. Souibgui, Y. Kessentini, and A. Fornés, “Enhance to read better: A multi-task adversarial network for handwritten document image enhancement,” *Pattern Recognition*, vol. 123, p. 108370, 2022.
- [4] M. Hradiš, J. Kotera, P. Zemcik, and F. Šroubek, “Convolutional neural networks for direct text deblurring,” in *Proceedings of BMVC*, vol. 10, no. 2, 2015.

- [5] B. Wang and C. L. P. Chen, "An effective background estimation method for shadows removal of document images," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 3611–3615.
- [6] M. A. Souibgui and Y. Kessentini, "De-gan: A conditional generative adversarial network for document enhancement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
- [8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.
- [10] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.
- [11] A. F. Biten, R. Litman, Y. Xie, S. Appalaraju, and R. Manmatha, "Latr: Layout-aware transformer for scene-text vqa," *arXiv preprint arXiv:2112.12494*, 2021.
- [12] A. C. Rouhou, M. Dhiaf, Y. Kessentini, and S. B. Salem, "Transformer-based approach for joint handwriting and named entity recognition in historical document," *Pattern Recognition Letters*, 2021.
- [13] V. De Bortoli, A. Desolneux, B. Galerne, and A. Leclaire, "Patch redundancy in images: A statistical testing framework and some applications," *SIAM Journal on Imaging Sciences*, vol. 12, no. 2, pp. 893–926, 2019.
- [14] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.
- [15] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.
- [16] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern recognition*, vol. 33, no. 2, pp. 225–236, 2000.
- [17] W. Xiong, J. Xu, Z. Xiong, J. Wang, and M. Liu, "Degraded historical document image binarization using local features and support vector machine (svm)," *Optik*, vol. 164, pp. 218–223, 2018.
- [18] R. Hedjam, M. Cheriet, and M. Kalacska, "Constrained energy maximization and self-referencing method for invisible ink detection from multispectral historical document images," in *2014 22nd International Conference on Pattern Recognition*. IEEE, 2014, pp. 3026–3031.
- [19] I. Pratikakis, K. Zagoris, G. Barlas, B. Gatos, "Icdar 2017 competition on document image binarization (dibco 2017)," in *2017 International Conference on Document Analysis and Recognition*. IEEE, 2017, pp. 1395—1403.
- [20] M. Z. Afzal, J. Pastor-Pellicer, F. Shafait, T. M. Breuel, A. Dengel, and M. Liwicki, "Document image binarization using lstm: A sequence learning approach," in *Proceedings of the 3rd international workshop on historical document imaging and processing*, 2015, pp. 79–84.
- [21] X.-J. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016, pp. 2810–2818.
- [22] K. G. Lore, A. Akintayo, and S. Sarkar, "Llnet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognition*, vol. 61, pp. 650–662, 2017.
- [23] J. Calvo-Zaragoza and A.-J. Gallego, "A selectional auto-encoder approach for document image binarization," *Pattern Recognition*, vol. 86, pp. 37–47, 2019.
- [24] Y. Akbari, S. Al-Maadeed, and K. Adam, "Binarization of degraded document images using convolutional neural networks and wavelet-based multichannel images," *IEEE Access*, vol. 8, pp. 153517–153534, 2020.
- [25] C. Tensmeyer and T. Martinez, "Document image binarization with fully convolutional neural networks," in *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, vol. 1. IEEE, 2017, pp. 99–104.
- [26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [27] J. Zhao, C. Shi, F. Jia, Y. Wang, and B. Xiao, "Document image binarization with cascaded generators of conditional generative adversarial networks," *Pattern Recognition*, vol. 96, p. 106968, 2019.
- [28] A. K. Bhunia, A. K. Bhunia, A. Sain, and P. P. Roy, "Improving document binarization via adversarial noise-texture augmentation," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 2721–2725.
- [29] M. O. Tamrin, M. El-Amine Ech-Cherif, and M. Cheriet, "A two-stage unsupervised deep learning framework for degradation removal in ancient documents," in *Pattern Recognition. ICPR International Workshops and Challenges*. Springer International Publishing, 2021, pp. 292–303.
- [30] M. A. Souibgui, Y. Kessentini, and A. Fornés, "A conditional gan based approach for distorted camera captured documents recovery," in *Mediterranean Conference on Pattern Recognition and Artificial Intelligence*. Springer, 2020.
- [31] Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, Y. Lu, D. Florencio, C. Zhang, W. Che *et al.*, "Layoutlmv2: Multi-modal pre-training for visually-rich document understanding," *arXiv preprint arXiv:2012.14740*, 2020.
- [32] S. Appalaraju, B. Jasani, B. U. Kota, Y. Xie, and R. Manmatha, "Docformer: End-to-end transformer for document understanding," *ICCV*, 2021.
- [33] P. Li, J. Gu, J. Kuen, V. I. Morariu, H. Zhao, R. Jain, V. Manjunatha, and H. Liu, "Selfdoc: Self-supervised document representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5652–5660.
- [34] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1833–1844.
- [35] H. Feng, Y. Wang, W. Zhou, J. Deng, and H. Li, "Doctr: Document image transformer for geometric unwarping and illumination correction," *arXiv preprint arXiv:2110.12942*, 2021.
- [36] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.
- [37] I. Pratikakis, K. Zagoris, P. Kadas, and B. Gatos, "Icfhr 2018 competition on handwritten document image binarization (h-dibco 2018)," in *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2018, pp. 489–493.
- [38] I. Pratikakis, B. Gatos, K. Ntirogiannis, "H-dibco 2010 - handwritten document image binarization competition," in *International Conference on Frontiers in Handwriting Recognition*. IEEE, 2010, pp. 727—732.
- [39] K. N. I. Pratikakis, B. Gatos, "Icdar 2011 document image binarization contest (dibco 2011)," in *2011 International Conference on Document Analysis and Recognition*, 2011, p. 1506–1510.
- [40] J.-C. Burie, M. Coustaty, S. Hadi, M. W. A. Kesiman, J.-M. Ogier, E. Paulus, K. Sok, I. M. G. Sunarya, and D. Valy, "Icfhr2016 competition on the analysis of handwritten text in images of balinese palm leaf manuscripts," in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2016, pp. 596–601.
- [41] Q.N. Vo, S.H. Kim, H.J. Yang, G. Lee, "Binarization of degraded document images based on hierarchical deep supervised network," *Pattern Recognition*, vol. 74, pp. 568—586, 2018.
- [42] I. Pratikakis, B. Gatos and K. Ntirogiannis, "ICFHR 2012 competition on handwritten document image binarization (H-DIBCO 2012)," in *Proceedings of the International Conference on Frontiers in Handwriting Recognition*. IEEE, 2012, pp. 817—822.