Text-DIAE: A Self-Supervised Degradation Invariant Autoencoder for Text Recognition and Document Enhancement

Mohamed Ali Souibgui*[1], Sanket Biswas*[1], Andres Mafla*[1], Ali Furkan Biten*[1], Alicia Fornés[1], Yousri Kessentini[2], Josep Lladós[1], Lluis Gomez[1], Dimosthenis Karatzas[1]

[1] Computer Vision Center, Universitat Autònoma de Barcelona, Barcelona, Spain
{msouibgui, sbiswas, amafla, abiten, afornes, josep, lgomez, dimos}@cvc.uab.es
[2] Digital Research Center of Sfax, SM@RTS Laboratory, Sfax, Tunisia
yousri.kessentini@crns.rnrt.tn

Abstract

In this paper, we propose a Text-Degradation Invariant Auto Encoder (Text-DIAE), a self-supervised model designed to tackle two tasks, text recognition (handwritten or scene-text) and document image enhancement. We start by employing a transformer-based architecture that incorporates three pretext tasks as learning objectives to be optimized during pre-training without the usage of labeled data. Each of the pretext objectives is specifically tailored for the final downstream tasks. We conduct several ablation experiments that confirm the design choice of the selected pretext tasks. Importantly, the proposed model does not exhibit limitations of previous state-of-the-art methods based on contrastive losses, while at the same time requiring *substantially* fewer data samples to converge. Finally, we demonstrate that our method surpasses the state-of-the-art in existing supervised and self-supervised settings in handwritten and scene text recognition and document image enhancement. Our code and trained models will be made publicly available at http://Upon_Acceptance.

1 Introduction

In recent times, self-supervised learning paradigms have gained a lot of attention due to its ability of benefiting from massive unlabeled data which is easily accessible from different sources. However, applying these approaches remain quite limited in the domains of optical character recognition (OCR), handwritten text recognition (HTR) and document image enhancement, which motivate us to tackle the problem in this study.

Common computer vision pipelines using self-supervised frameworks employ a pretext-task (e.g. relative position prediction of patches (Doersch, Gupta, and Efros 2015), contrastive views (Chen et al. 2020a), image inpainting (Pathak et al. 2016), etc.) to learn visual representations for solving down-stream tasks like classification, object detection and so on. Current self-supervised paradigms (Caron et al. 2021, 2020; Chen et al. 2020a; Chen, Xie, and He 2021) have adapted transformers (Vaswani et al. 2017) to learn visual representations from unlabeled images which are semantically meaningful. More recently, generative self-supervised approaches (He et al. 2021; Bao, Dong, and Wei 2021; Dong et al. 2021) using auto-encoders have been used to learn representations in the feature space through image patches and visual tokens.

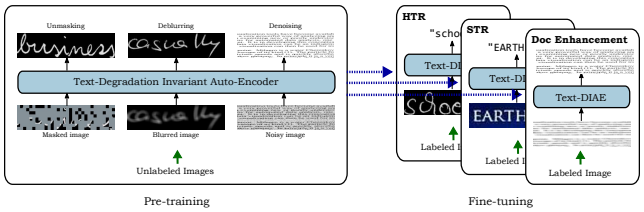


Figure 1: **Text-Degradation Invariant Auto-Encoder (Text-DIAE),** we employ image reconstruction pretext tasks at pre-training. Masking, blurring and adding noise are employed to learn richer representations.

Closely related to our work, some contributions in visual representation learning were addressing text recognition (HTR) (Aberdam et al. 2021; Bhunia et al. 2021; Liu et al. 2022) and Scene-Text Recognition (STR) (Aberdam et al. 2021; Zhang et al. 2022)) and image enhancement (Liang et al. 2022). Despite the performance gains, there are some drawbacks of such models: (1) independent sequences of tokens are treated as single data points, which can cause misalignment of similar sequences among a batch, (2) considerable batch size requirements to define negative contrastive pairs, (3) considerably slow convergence rates.

For humans, reading text in noisy scenarios is possible because of the ability or reconstructing the degraded regions and predicting the missing/blurry content (Howard et al. 1998; Dehaene 2014). Incorporating such an ability in a model could immensely help in restoration, recognition and understanding of characters and symbols, considering that text carries rich linguistic information that allow humans to reason explicitly according to context. In order to endow this human-specific skill to our models, we present in this paper a new self-supervised framework called Text-Degradation Invariant Auto-Encoders (Text-DIAE) inspired

*These authors contributed equally.

by the principle of denoising autoencoders (Vincent et al. 2008), as depicted in Figure 1. Our model focuses on exploring the dynamics of learning representations under different degradation scenarios. Specifically, we propose the usage of a robust self-supervised auto-encoder along with customized pretext tasks (masking, blur and background noise) that are designed to specifically tackle two different downstream tasks: text recognition (HTR and STR) and document image enhancement (document binarization, document deblurring). As a consequence, the choice of the proxy tasks have been realized to learn useful representations for solving these specific downstream tasks using unlabeled data.

The benefits of employing such approach are: we do not define sequences at the feature level. Rather, by employing a transformer-based (Vaswani et al. 2017) approach, similar to BERT (Devlin et al. 2018) we utilize the self-attention layers to attend among patches which does not require big batches of negative samples. Also, the combination of these pre-training tasks result in a significantly faster convergence compared to previous approaches. The resulting representations are evaluated by a scenario that resembles the linear probing evaluation often used in self-supervision (Kolesnikov, Zhai, and Beyer 2019; Zhang, Isola, and Efros 2016) and follows the scheme of (Aberdam et al. 2021) in text recognition task. By this assessment, we find that our method outperforms previous self and semi supervised pipelines. Furthermore, by employing Text-DIAE, we achieve state-of-the-art in handwritten text recognition and document image enhancement, while outperforming scene text recognition under self-supervision settings. The essential findings and novelties of our work are based on the following interesting deductions:

- The impact and combination of pretext tasks depends on the downstream task.
- The closer the association between a pretext task and a downstream task, the better is the model performance.
- By employing Text-DIAE, we achieve faster convergence and use order of magnitude lesser data during pre-training than the contrastive-learning based approaches.

To add on top of this, this is the first work to our knowledge that investigates different self-supervised pretext tasks for multiple significant downstream tasks in text recognition (HTR-word level, STR) and document image enhancement (document binarization, deblurring) while achieving state-of-the-art performance with 43 and 45 times lesser data for HTR and STR, respectively.

2 Related Work

Self-Supervised Learning. Due to extensive efforts on labeled data requirements of supervised models, this learning paradigm emerges as a way of exploiting the structured information contained in data itself. Self-Supervised learning aims to obtain rich representations of an input modality by designing pretext tasks that are used as auxiliary signals that are informative for a given downstream task. Initial approaches relied on auto-encoders (Vincent et al. 2008) trained to remove artificially added noise from an image. Later, several approaches introduced other pretext tasks that provide rich signals to train a network as a feature extractor. Some pretext tasks employed were image colorization (Zhang, Isola, and Efros 2016), jigsaw puzzle solving (Noroozi and Favaro 2016), patch ordering (Doersch, Gupta, and Efros 2015), rotation prediction (Gidaris, Singh, and Komodakis 2018) among others. Recent approaches rely on extensive image augmentation to maximize the agreement among paired samples and contrast with all possible negative samples (Chen et al. 2020a,b; He et al. 2020; Zbontar et al. 2021; Caron et al. 2020, 2021).

More recently, generative approaches like Masked Autoencoders (MAE) (He et al. 2021) are introduced to predict a masked latent representation of patches. Similar ideas have been also explored in other recent works like BEiT (Bao, Dong, and Wei 2021) and PeCo (Dong et al. 2021) which adopt a discrete variational autoencoder (VAE) to generate discrete visual tokens from the original image. Motivated by these works, we expand this generative learning framework to tackle text recognition and document enhancement tasks.

Text Recognition. Ample research in text recognition has been conducted, resulting in handwritten (HTR) (Sonkusare and Sahu 2016; Memon et al. 2020) and scene-text (STR) (Shi, Bai, and Yao 2016; Long, He, and Yao 2021; Chen et al. 2021) recognition pipelines. Most common approaches that tackle text recognition are using supervised methodologies that employ an encoder-decoder mechanism (Cheng et al. 2017; Shi, Bai, and Yao 2016; Shi et al. 2016; Litman et al. 2020; Kang et al. 2020a) based on a Connectionist Temporal Classification (CTC) (Graves et al. 2006) network or an Attention-based (Cheng et al. 2017; Shi et al. 2016) decoder. Recently, approaches that focus on semi-supervised and self-supervised learning have been explored (Souibgui et al. 2021) with domain adaptation techniques on STR (Kang et al. 2020b) and HTR (Zhang et al. 2019). Under the unsupervised paradigm, (Gupta, Vedaldi, and Zisserman 2018) formulate text recognition as a task to align the conditional distribution of strings predicted with lexically correct strings sampled from a text database. Closely related to our work, (Aberdam et al. 2021) proposes a self-supervised sequence-to-sequence model that separates consecutive text features to be later used in a contrastive loss similar to (Chen et al. 2020a). Analogously, (Zhang et al. 2022) and (Liu et al. 2022) improve the features obtained from a contrastive loss by concatenating characters and by perceiving spatial strokes respectively. Nevertheless, these methods require large batches, and rely on a sequential definition of features that can produce misaligned characters or n-grams contained in different words.

Document Image Enhancement. Many approaches have been proposed to address the enhancement of documents (both handwritten and machine-printed) which suffer several kinds of artefacts/defects such as bleed-through, show-through, faint characters, contrast variations and so on. Adaptive thresholding based on sliding-window operations by (Sauvola and Pietikäinen 2000) formulated a strong handcrafted baseline for binarization task. The work from (Calvo-Zaragoza and Gallego 2019; Kang, Iwana, and Uchida 2021) map images from the degraded domain to the enhanced one using end-to-end CNN-based autoen-

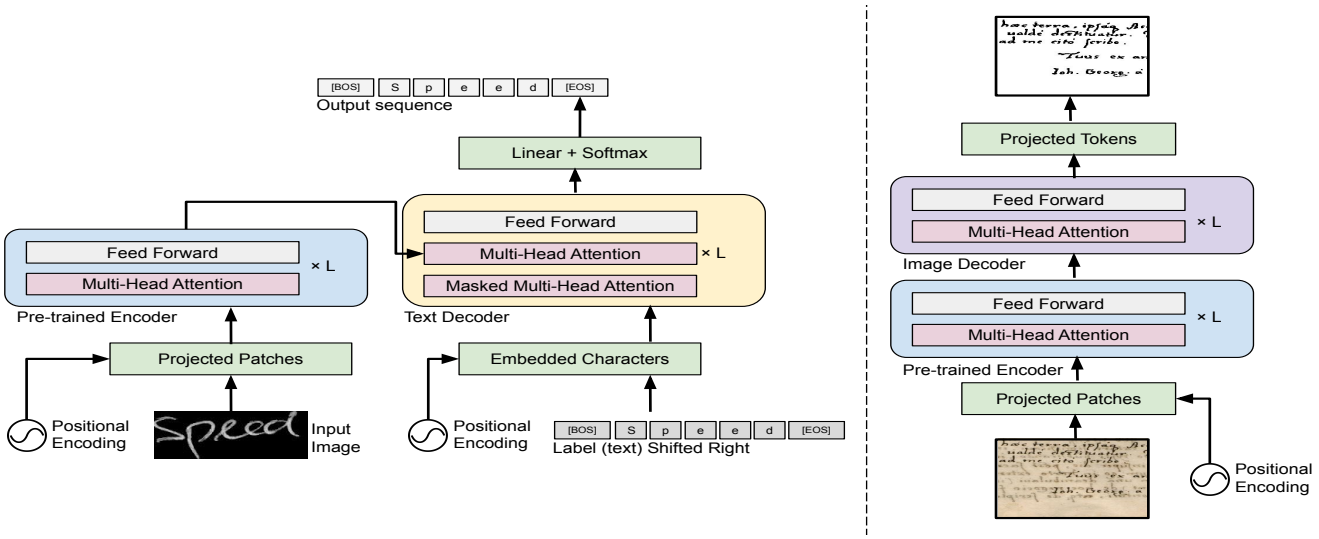


Figure 2: **Fine-tuning pipeline**. We start from a pretrained encoder as initial weights to solve a specific downstream task. Explicit decoders are used for text recognition (left) and document image enhancement (right).

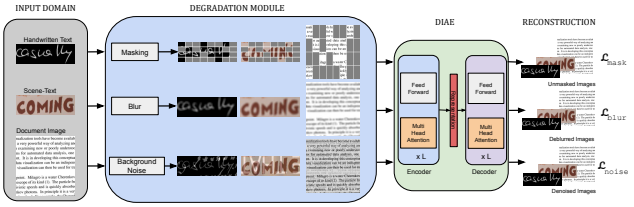


Figure 3: **Pre-training pipeline.** Text-DIAE aims to learn degradation invariant representations. These are later used to reconstruct the input image with a specific learning objective for each degradation type.

coders. Other techniques (Zhao et al. 2019; Souibgui and Kessentini 2020; Souibgui, Kessentini, and Fornés 2020; Jemni et al. 2022) used conditional-Generative Adversarial Network (c-GAN) based approaches to design a generator which produces the enhanced version of the document while the discriminator assesses the quality of binarization. Lately, an end-to-end ViT autoencoder was proposed in (Souibgui et al. 2022) to capture high-level global features using self-attention for binarizing degraded documents. Regarding document deblurring, a benchmark was formulated by (Hradiš et al. 2015) where a CNN were trained to reconstruct enhanced high-quality images from blurry inputs that consist of a combination of camera-driven motion blurred and de-focused images of text documents. Lately, (Souibgui and Kessentini 2020) improved the baseline performance using the similar c-GAN based approach aimed in a binarization task.

3 Method

In this section, we present our proposed method for text image recognition and enhancement by describing its building blocks. Our approach uses two steps: a pre-training stage to learn useful representations from unlabeled data, and a supervised fine-tuning phase for the desired downstream task.

3.1 Pre-Training Module

The overall pre-training pipeline of Text-DIAE is shown in Fig. 3. For each task, given an unlabeled image I (eg. a cropped handwritten text, cropped scene text or a scanned document image), we use a function ϕ to map I to a degraded form. The function ϕ takes as parameters the original image I and the degradation type $\mathcal{T} \in \{mask, blur, noise\}$ where we denote a degraded image by $I_d = \phi(I, \mathcal{T})$.

Our model is composed of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$ with learnable parameters $\theta_\mathcal{E}$, $\theta_\mathcal{D}$ respectively. The pre-training pipeline trains an encoder function $\mathcal{E}$ that maps the degraded image I_d to a latent representation $z_\mathcal{T}$ in a multi task fashion (unmasking, deblurring and denoising) and then learning a decoder $\mathcal{D}$ to reconstruct the original image I from the representation $z_\mathcal{T}$:

$$\begin{aligned} z_\mathcal{T} &= \mathcal{E}(\phi(I, \mathcal{T}); \theta_E) \\ I_r &= \mathcal{D}(z_\mathcal{T}; \theta_D) \end{aligned} \tag{1}$$

The learned visual representations from the latent subspace should be invariant to the applied degradation $\mathcal{T}$.

Encoder. The encoder architecture consists of a vanilla ViT (Dosovitskiy et al. 2021) backbone. Given an input image I_d, it is first split into a set of N patches, $I_d^p = \{I_d^{p_1}, I_d^{p_2}, ..., I_d^{p_N}\}$. Then, these patches are embedded with a trainable linear projection layer E. Text-DIAE uses a distinct linear projection layer for every defined pre-text task. These tokens are later concatenated with their 2-D positional information embedded with E_{pos} and fed to L transformer blocks to map these tokens to the encoded latent representation z_l. These blocks are composed of L layers of Multi-head Self-Attention (MSA) and a feedforward Multi-Layered Perceptron (MLP) as depicted in Figure 3. Each of these blocks

Table 1: **Representation quality.** We evaluate the encoder capability of learning visual representations. This scenario is analogous as the linear probing in self-supervised models. We train a decoder with labelled data on top of a frozen encoder pre-trained on the proposed degradation. The column *Seen* refers to the number of samples in millions seen during pre-training. Word prediction in terms of Accuracy (Acc) and single edit distance (ED1) in handwritten and text recognition.

Method	Encoder	Decoder	Handwritten Text						Scene-Text					
			IAM			CVL			IIIT5K			IC13		
			Acc	ED1	Seen	Acc	ED1	Seen	Acc	ED1	Seen	Acc	ED1	Seen
simCLR (Chen et al. 2020a)	CNN	CTC	4.0	16.0	205.8	1.8	11.1	205.8	0.3	3.1	409.6	0.3	5.0	409.6
seqCLR (Aberdam et al. 2021)			39.7	63.3	205.8	66.7	77.0	205.8	35.7	62.0	409.6	43.5	67.9	409.6
PerSec (Liu et al. 2022)			–	–	–	–	–	–	37.9	–	–	46.4	–	–
PerSec (Liu et al. 2022)	ViT		–	–	–	–	–	–	38.4	–	–	46.7	–	–
simCLR (Chen et al. 2020a)	CNN	Attn.	16.0	21.2	205.8	26.7	30.6	205.8	2.4	3.6	409.6	3.1	4.9	409.6
seqCLR (Aberdam et al. 2021)			51.9	65.0	205.8	74.5	77.1	205.8	49.2	68.6	409.6	59.3	77.1	409.6
PerSec (Liu et al. 2022)			–	–	–	–	–	–	50.7	–	–	61.1	–	–
PerSec (Liu et al. 2022)	ViT		–	–	–	–	–	–	52.3	–	–	62.3	–	–
Ours	ViT	Transf.	**71.0**	**82.1**	**4.7**	**78.1**	**81.5**	**1.2**	**77.1**	**87.8**	**9.1**	**92.6**	**95.6**	**18.2**

are preceded by a LayerNorm (LN) (Ba, Kiros, and Hinton 2016) and followed by a residual connection:

$$\begin{aligned} z_0 &= E(I_d^p) + E_{pos} \\ z'_l &= \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1},\ l = 1, \ldots \text{L} \\ z_l &= \text{MLP}(\text{LN}(z'_l)) + z'_l,\ l = 1, \ldots \text{L} \\ z_{\mathcal{T}} &= LN(z_L) \end{aligned} \tag{2}$$

Decoder. The decoder composed of transformer blocks following the same structure and number of layers as the encoder. The decoder input is the output of encoder $z_{\mathcal{T}}$. The output of the decoder is a set of vectors $I_r = \{I_r^{p_1}, I_r^{p_2}, ..., I_r^{p_N}\}$ where each of which corresponds to a flattened patch in the predicted (reconstructed) image. Same as before, a distinct linear layer is used for each pre-text task.

$$\begin{aligned} z'_l &= \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1}\ ,\ l = 1, \ldots \text{L} \\ z_l &= \text{MLP}(\text{LN}(z'_l)) + z'_l\ ,\ l = 1, \ldots \text{L} \\ I_r &= \text{Linear}(z_L) \end{aligned} \tag{3}$$

3.2 Fine-Tuning

Our fine tuning process is illustrated in Fig. 2 where we perform two different downstream tasks; text recognition and document image enhancement.

Text Recognition. Text recognition aims to transform an image into the machine encoded form, i.e., sequence of characters. Let I be a cropped text image and $C = \{c_1, c_2, ..., c_N\}$ its ground truth label which corresponds to a sequence of characters, where N is the length of the text. The training is done by passing I to an encoder function $\mathcal{E}$ to produce a latent representation z. Then, z is later fed to a decoder function $\mathcal{D}'$ to produce a sequence of characters $C_p = \{c_{p_1}, c_{p_2}, ..., c_{p_N}\}$ that should match the ground truth label sequence.

We initialize the encoder with the pre-trained weights $\theta_{\mathcal{E}}$ while we employ a sequential transformer decoder (Vaswani et al. 2017) as seen in Fig. 2-Left. The decoder is initialized randomly and composed of L transformer blocks of MSA, MLP and Masked-MSA layers preceded by LN layers, and followed by a residual connection. The output of the decoder is a sequence of characters where at each time step t, the predicted character is formed by attending to the representation z and previous character embeddings until $t-1$.

Document Image Enhancement. Document enhancement consists of mapping a degraded document into a clean form. Let I_d be a degraded image and I_c its clean version, then the goal is to learn an encoder function $\mathcal{E}$ that maps I_d to a representation z with the same way as in Eqn 2. $\mathcal{E}$ weights are initialized from the pre-training stage. The decoder $\mathcal{D}''$ generates the clean image I_c from z as in Eqn 3.

3.3 Learning Objectives

Our model makes use of different sets of losses for each phase. During pre-training, we use three different losses. Each one is dedicated to a particular pre-text task: $\mathcal{L}_{mask}$, $\mathcal{L}_{blur}$ and $\mathcal{L}_{noise}$. Each of these losses is a mean squared error (MSE) between the reconstructed image I_r (from the masked, blurred or noisy image) and its ground-truth version I_{gt}. Thus, the overall loss for our pre-training stage is:

$$\mathcal{L}_{pt} = \lambda_1 \mathcal{L}_m\left(\text{I}_r, \text{I}_{gt}\right) + \lambda_2 \mathcal{L}_b\left(\text{I}_r, \text{I}_{gt}\right) + \lambda_3 \mathcal{L}_n\left(\text{I}_r, \text{I}_{gt}\right) \tag{4}$$

During our experimentation, the best results were obtained with setting $\lambda_1 = \lambda_2 = \lambda_3 = 1$.

While fine-tuning on text recognition, we use a cross-entropy loss between the predicted sequence of characters C_p and C. For document image enhancement fine tuning, we used an MSE loss between the cleaned image I_c and I.

4 Experiments

In this section we describe the studied scenarios and experiments performed for text recognition and document enhancement respectively. We ask the reader to refer to the supplementary material for specific implementation details.

4.1 Text Recognition

Evaluating Representations. In order to evaluate the quality of the learned representations, and extending commonly

Method	Encoder	Decoder	Handwritten Text						Scene-Text	
			IAM			CVL			IIIT5K	IC13
			5%	10%	100%	5%	10%	100%	100%	100%
Supervised (Aberdam et al. 2021)	CNN	CTC	21.4	33.6	75.2	48.7	63.6	75.6	76.1	84.3
simCLR (Chen et al. 2020a)			15.4	21.8	65.0	52.1	62.0	74.1	69.1	79.4
seqCLR (Aberdam et al. 2021)			31.2	44.9	76.7	66.0	71.0	77.0	80.9	86.3
PerSec (Liu et al. 2022)			–	–	77.9	–	–	78.1	82.2	87.9
PerSec (Liu et al. 2022)	ViT		–	–	78.0	–	–	78.8	83.7	89.7
Supervised (Aberdam et al. 2021)	CNN	Attn.	25.7	42.5	77.8	64.0	72.1	77.2	83.8	88.1
simCLR (Chen et al. 2020a)			22.7	32.2	70.7	59.0	65.6	75.7	77.8	84.9
seqCLR (Aberdam et al. 2021)			40.3	52.3	79.9	**73.1**	**74.8**	77.8	82.9	87.9
PerSec (Liu et al. 2022)			–	–	80.8	–	–	80.2	84.2	88.9
PerSec (Liu et al. 2022)	ViT		–	–	**81.8**	–	–	80.8	85.2	89.2
Supervised (Ours)	ViT	Transf.	22.8	25.3	71.7	17.9	19.8	71.9	75.7	91.9
Ours			**49.6**	**58.7**	80.0	47.9	68.5	**87.3**	**86.1**	**92.0**

Table 2: **Semi-supervised results.** Accuracy obtained by fine-tuning a pre-trained model with varying percentages of the labeled dataset. Under this setting, we back-propagate the gradients through the specific decoder and the pre-trained encoder.

Table 3: **SOTA results.** Quantitative evaluation with state-of-the-art methods on the IAM word level dataset.

Method	CER↓	WER↓	Avg.
Bluche et al. (Bluche 2015)	**7.3**	24.7	16.00
Bluche et al. (Bluche 2016)	7.9	24.6	16.25
Sueiras et al. (Sueiras et al. 2018)	8.8	23.8	16.30
ScrabbleGAN (Fogel et al. 2020)	-	23.6	-
SSDAN (Zhang et al. 2019)	8.5	22.2	15.35
SeqCLR (Aberdam et al. 2021)	9.5	20.1	14.80
PerSec (Liu et al. 2022)	-	18.2	-
Ours	9.3	**20.0**	**14.65**

used linear-probing settings (Zhang, Isola, and Efros 2016), we employ a similar approach as introduced by (Aberdam et al. 2021). As a first step, the encoder is pre-trained with unlabeled data as described in Section 3.1. After that, the encoder's weights are frozen and a new decoder is trained on top of it with all the labeled data. The decoder, as we detailed above, generates the predicted characters in a time-step manner. Since the encoder remains frozen, this scenario is a good proxy that represents the expressivity of the learned visual representations. To this end, Table 1 shows the results of our proposed approach. We compare among self-supervised methods specifically designed for the text recognition task.

Better performance. As it can be seen from Table 1, the seqCLR method presented by (Aberdam et al. 2021) improves significantly a self-supervised baseline inspired by SimCLR (Chen et al. 2020a). In the recently released approach PerSec by (Liu et al. 2022), they slightly improve over the seqCLR. It is evident that our Text-DIAE model *greatly* outperforms all the aforementioned state-of-the-art approaches regarding the representation quality obtained, both in handwritten and scene-text. The improvements in term of the accuracy in a handwritten text dataset, IAM, is close to **+20 points**. Moreover, a bigger improvement gap is obtained when evaluating scene-text. An average gain of **+30 points** is accomplished in IIIT5K and ICDAR13, proving the generalization of our method to different domains. In our model, the great expressivity of features achieved by the encoder is mainly due to two factors. Firstly, by masking image patches, the encoder learns a strong unigram character distribution (refer to Figure 4), which is not leveraged in previous methods. Secondly, by distorting and recovering the image, we make the model learn richer representations to detect and recover the text into a clean and readable state. Thus, the model is learning the most valuable features that lead to the best recognition performance.

Faster convergence. One of the most important outcomes by employing our method, is that a **paramount** improvement in convergence is achieved during pre-training. Table 1 shows this effect under the column labeled as "Seen". It depicts the total number of seen samples that each model requires during the pre-training stage. It is worth highlighting that during pre-training the encoder of our model requires **43** and **166** times lesser data in IAM and CVL respectively when compared to the seqCLR and simCLR. In scene-text, our model employs only 18.2M samples to yield powerful representations compared to the 409M samples required by previous self-supervised approaches.

Fine-Tuning. In this stage, we evaluate our model considering a semi-supervised setting where the obtained results are depicted in Table 2. Here we use the self-supervised pre-trained encoder as a backbone and train a transformer-based decoder from scratch that predicts the characters in a sequential manner, as illustrated in Fig. 2-Left. In this scenario, the gradients are back-propagated not only to the decoder but also to the encoder. Following the previous work (Aberdam et al. 2021), we use 5% and 10% of the labeled dataset by randomly selecting the training samples. As suggested in (Chen et al. 2020a) we perform fine-tuning on all the labeled dataset. In order to compare with (Aberdam et al. 2021) and since scene-text dataset is synthetic, we evaluate with the complete labeled dataset.

Higher performance in fine-tuning settings. Our model exploits data in a more efficient manner than previous self-supervised methods in fine-tuning setting. We infer that the set of degradations proposed yields rich signals, helping the encoder to adapt to the downstream task more efficiently. Our model achieves state-of-the-art in all scenarios when all

Table 4: **Ablations of the pre-training objectives.** Results in handwritten and scene-text recognition obtained by each pretext task. The performance is measured in terms of Word and Character error rates (WER and CER).

$\mathcal{L}_{mask}$	$\mathcal{L}_{blur}$	$\mathcal{L}_{noise}$	IAM CER↓	IAM WER↓	IAM Avg.	IC13 CER↓	IC13 WER↓	IC13 Avg.
✓	✗	✗	**9.3**	**20.0**	**14.65**	4.5	8.0	6.25
✓	✓	✗	12.3	24.8	18.5	**4.2**	**8.0**	**6.10**
✓	✗	✓	11.1	23.3	17.2	4.8	8.6	6.70
✓	✓	✓	11.4	23.8	17.6	5.1	9.3	7.20

Figure 4: **Qualitative results of pre-training samples.** The left refers to handwritten text, while scene-text is depicted on the right. On each scenario, from left to right, the original, masked and reconstructed images are depicted.

the labeled datasets are used except in IAM where the PerSec is slightly better. Under semi-supervised settings, our model performs better at the IAM dataset when employing 5% and 10% of the labels than simCLR and seqCLR. Since CVL contains substantially fewer data samples than IAM, SeqCLR still outperforms our approach in the CVL dataset. However, while employing the full labels of CVL, Text-DIAE outperforms all the methods by a large margin.

More efficient than a supervised baseline. From table 2, we can also notice the superiority of pre-training our architecture compared to a fully supervised model starting from scratch. This suggest that the self-supervised pre-training of such transformer-based architectures is essential to obtain better results, and beneficial especially in small labeled datasets scenarios, since the unlabeled data is generally easier to obtain for a self-supervised pre-training.

The effect of fine-tuning after pre-training. By proposing the degradation invariant optimization at pre-training, our model achieves a significant gain in recognition on handwritten text datasets. An average of 10 points of accuracy are gained after fine-tuning (refer to Table 1 and 2). Finally, it is important to note that our model reaches state-of-the-art in the handwritten text recognition task, even comparing to specifically designed supervised approaches. The results on the IAM dataset are shown in Table 3, which measures the performance of a model in terms of word and character error rate, WER and CER respectively.

Ablation Studies. The results of experimentation regarding the effect of each degradation as pretext task at pre-training is given in Table 4. Firstly, among the three proposed degradations, masking is the most crucial to be applied in both tasks, handwritten and scene text recognition. When an input word is masked, and in order to properly reconstruct it,

Table 5: **SOTA results:** Quantitative evaluation with state-of-the-art methods on the deblurring dataset.

Method	PSNR
CNN-Baseline (Hradiš et al. 2015)	19.36
Pix2Pix-HD (Wang et al. 2018)	19.89
DE-GAN (Souibgui and Kessentini 2020)	20.37
DocEnTr (Souibgui et al. 2022)	21.28
Ours	**23.58**

the model has to learn a character level distribution. This by itself provides with a strong prior compared to denoising or deblurring an image. Additionally, adding blur in scene-text imagery improves the representations learned by the model shown by the results. Lastly, adding noise does not result in an improvement in text recognition tasks. However, as it is shown in the next section, the combination of the 3 degradation produce a richer encoder in document enhancement. Therefore, we can safely assume that each degradation has a task-dependent impact on the representations learned depending on the similarity of them when compared to the final downstream task and input data distribution.

Qualitative Results. In Figure 4 we show the reconstructed images at pre-training stage for handwritten and scene-text samples. It is important to note the complexity of the reconstruction task even for humans. Even though high masking percentages are employed (75%), our model learns to properly adapt to handwritten styles and fonts found in scene-text. As can be appreciated, although sometimes our model's reconstruction does not match with the ground truth images, it can still reconstruct the most probable and plausible English words (e.g. see "school" vs "sand" in 4th row in handwritten examples). Another interesting outcome is also noticed for scene-text example where "xperia" is reconstructed correctly while the last character "a" is selected from another font, demonstrating the model's capability. Minor reconstruction errors are found such as that the model eventually learns to overcome at fine-tuning stage.

4.2 Document Image Enhancement

Performance Analysis on Binarization. As shown in Table 6, the Text-DIAE outperforms the previous state-of-the-art approaches on majority of the standard metrics for document binarization task. Specifically, the quantitative comparison of results demonstrate that Text-DIAE achieves an optimal gain in PSNR, FM, F_{ps} and DRD performance surpassing the all previous arts. The largest performance improvement is obtained over the H-DIBCO 2012 while the least performance gain is obtained in the H-DIBCO 2018. One of the major concerns which degraded historical documents face is the show-through effect, which appears when ink impressions from one side of the document start appearing on the other side, making it almost impossible to read as shown in Appendix. The enhanced Text-DIAE output illustrates that it not only resolves the show-through but also sharpens and smoothens the edges of the foreground text approximately to the ground-truth level.

Table 6: **SOTA results.** Comparison of the proposed Text-DIAE compared to previous state-of-the-art approaches on the different DIBCO and H-DIBCO Benchmarks

Method	DIBCO Benchmarks															
	2011				2012				2017				2018			
	PSNR↑	FM↑	F_{ps}↑	DRD↓	PSNR↑	FM↑	F_{ps}↑	DRD↓	PSNR↑	FM↑	F_{ps}↑	DRD↓	PSNR↑	FM↑	F_{ps}↑	DRD↓
(Sauvola and Pietikäinen 2000)	15.60	82.10	-	8.50	16.71	82.89	87.95	6.59	14.25	77.11	84.1	8.85	13.78	67.81	74.08	17.69
(Kang, Iwana, and Uchida 2021)	19.90	**95.50**	-	1.80	21.37	95.16	96.44	1.13	15.85	91.57	93.55	2.92	19.39	89.71	91.62	**2.51**
(Zhao et al. 2019)	20.30	93.80	-	1.80	21.91	94.96	96.15	1.55	17.83	90.73	92.58	3.58	18.37	87.73	90.60	4.58
(Souibgui et al. 2022)	20.81	94.37	96.15	1.63	22.29	95.31	96.29	1.60	19.11	92.53	95.15	2.37	19.46	90.59	93.97	3.35
Ours	**21.29**	95.01	**96.86**	**1.48**	**23.66**	**96.52**	**97.04**	**1.10**	**19.64**	**93.84**	**95.71**	**1.93**	**19.95**	**91.32**	**94.44**	3.21

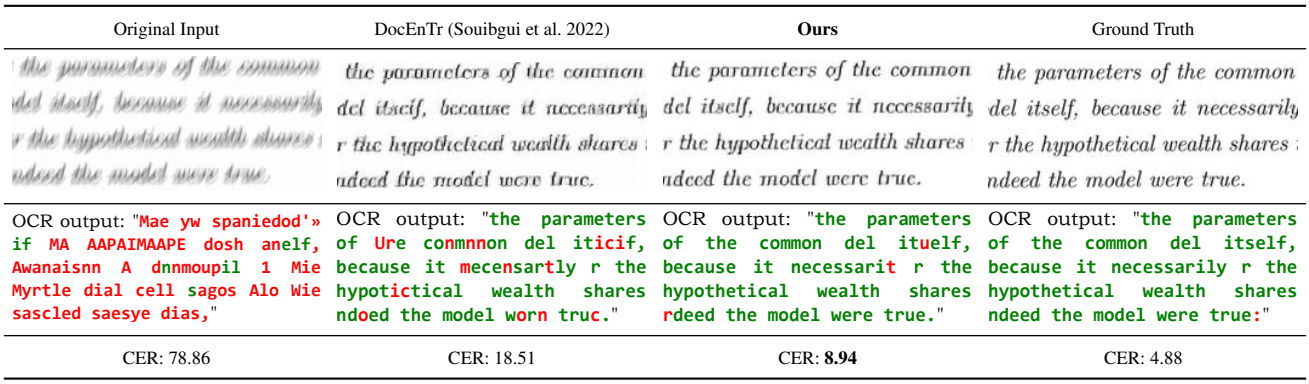


Figure 5: **Qualitative results of deblurred samples.** The document image on the left refers to the originally captured blurred image, followed by the ground-truth, and the deblurred results from the DocEnTr and our Text-DIAE model towards right. The correctly predicted OCR output is shown in "Green" font while the inaccurate ones are depicted in "Red" and recognition performance in terms of CER.

$\mathcal{L}_{mask}$	$\mathcal{L}_{blur}$	$\mathcal{L}_{noise}$	PSNR
✗	✗	✗	18.75
✓	✗	✗	19.65
✗	✓	✗	18.98
✗	✗	✓	19.82
✗	✓	✓	19.34
✓	✗	✓	19.45
✓	✓	✓	**19.95**

Table 7: **Ablations of the degradations as pre-training objectives.** Results in document image binarization on DIBCO 2018 obtained by each pretext task in terms of PSNR.

Performance Analysis on Deblurring. In Table 5 we show a quantitative comparison and superiority of Text-DIAE over supervised techniques (Hradiš et al. 2015; Wang et al. 2018; Souibgui and Kessentini 2020; Souibgui et al. 2022) on the document deblurring benchmark. A substantial gain in PSNR by **+2 points** on a **logarithmic** scale is obtained over DocEnTr (Souibgui et al. 2022), which signifies the greater quality of deblurred images generated by Text-DIAE. There are two different kinds of blurring which appear in documents: motion blur owing to the sudden rapid camera movement and out-of-focus blur which emerges when light fails to converge in the image. In Fig. 5, we show an interesting qualitative case study of a motion blurred document image. We assess the performance of deblurring by running the Tesseract-OCR engine (Smith 2007) over the blurred, ground-truth, DocEnTr prediction and the Text-DIAE output. Qualitative results show that Text-DIAE significantly decreases the CER, showing vast improvement in OCR performance as depicted in green font.

Ablation Studies. We also showcase an interesting ablation on the task of document image binarization for the challenging DIBCO 2018 benchmark. From Table 7, we infer that any pre-training task is beneficial while the denoising task is the most crucial to be applied when each pre-text task is applied separately. The aforementioned result can be attributed to the fact that denoising is much closer to the downstream binarization task. Also, it demonstrates that Text-DIAE performs the best for document enhancement tasks when the model learns all the possible degradation (masking, blurring and adding noise) together.

5 Conclusion

This work demonstrates the capability of learning richer representations through pretext degradation tasks. Self-supervised learning can immensely boost the performance of text recognition and document image enhancement without any requirement of labeled data. Notably, we show that Text-DIAE does not share the limitations of contrastive or sequential approaches and is more effective at learning rich representations while seeing *significantly* less data points. Extensive experimentation during fine-tuning demonstrate that Text-DIAE surpasses previous supervised and self-supervised state-of-the-art in handwritten text recognition and document image enhancement, while outperforming previous self-supervised approaches in scene-text recognition. We hypothesize that Text-DIAE performs complex variable reconstructions during pre-training, which helps to

learn meaningful visual concepts from the latent representation space. We also provide the community the following insights to work on : 1) Designing new pretext tasks that are similar to downstream tasks. 2) The effect/trade-off of combination of various pretext tasks on the downstream tasks. 3) A need for a holistic approach to combine all the tasks into a single model.

Acknowledgments

This work has been partially supported by the Swedish Research Council (grant 2018-06074, DECRYPT), the Spanish projects RTI2018-095645-B-C21, CERCA Program / Generalitat de Catalunya, the FCT-19-15244, the Catalan projects 2017-SGR-1783, PhD Scholarships from AGAUR (2021FIB-10010) and (2019-FIB01233), and from UAB (B18P0073). DocPRESERV project (Swedish STINT grant).

References

Aberdam, A.; Litman, R.; Tsiper, S.; Anschel, O.; Slossberg, R.; Mazor, S.; Manmatha, R.; and Perona, P. 2021. Sequence-to-sequence contrastive learning for text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15302–15312.

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bao, H.; Dong, L.; and Wei, F. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Bhunia, A. K.; Chowdhury, P. N.; Yang, Y.; Hospedales, T. M.; Xiang, T.; and Song, Y.-Z. 2021. Vectorization and rasterization: Self-supervised learning for sketch and handwriting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5672–5681.

Bluche, T. 2015. *Deep neural networks for large vocabulary handwritten text recognition*. Ph.D. thesis, Paris 11.

Bluche, T. 2016. Joint line segmentation and transcription for end-to-end handwritten paragraph recognition. *Advances in neural information processing systems*, 29.

Burie, J.-C.; Coustaty, M.; Hadi, S.; Kesiman, M. W. A.; Ogier, J.-M.; Paulus, E.; Sok, K.; Sunarya, I. M. G.; and Valy, D. 2016. 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)*, 596–601. IEEE.

Calvo-Zaragoza, J.; and Gallego, A.-J. 2019. A selectional auto-encoder approach for document image binarization. *Pattern Recognition*, 86: 37–47.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33: 9912–9924.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9650–9660.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; and Hinton, G. E. 2020b. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33: 22243–22255.

Chen, X.; Jin, L.; Zhu, Y.; Luo, C.; and Wang, T. 2021. Text recognition in the wild: A survey. *ACM Computing Surveys (CSUR)*, 54(2): 1–35.

Chen, X.; Xie, S.; and He, K. 2021. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9640–9649.

Cheng, Z.; Bai, F.; Xu, Y.; Zheng, G.; Pu, S.; and Zhou, S. 2017. Focusing attention: Towards accurate text recognition in natural images. In *Proceedings of the IEEE international conference on computer vision*, 5076–5084.

Dehaene, S. 2014. *Consciousness and the brain: Deciphering how the brain codes our thoughts*. Penguin.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Doersch, C.; Gupta, A.; and Efros, A. A. 2015. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, 1422–1430.

Dong, X.; Bao, J.; Zhang, T.; Chen, D.; Zhang, W.; Yuan, L.; Chen, D.; Wen, F.; and Yu, N. 2021. PeCo: Perceptual Codebook for BERT Pre-training of Vision Transformers. *arXiv preprint arXiv:2111.12710*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Fogel, S.; Averbuch-Elor, H.; Cohen, S.; Mazor, S.; and Litman, R. 2020. Scrabblegan: Semi-supervised varying length handwritten text generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4324–4333.

Gidaris, S.; Singh, P.; and Komodakis, N. 2018. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 249–256. JMLR Workshop and Conference Proceedings.

Gómez, L.; Mafla, A.; Rusinol, M.; and Karatzas, D. 2018. Single shot scene text retrieval. In *ECCV*.

Graves, A.; Fernández, S.; Gomez, F.; and Schmidhuber, J. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks.

In *Proceedings of the 23rd international conference on Machine learning*, 369–376.

Gupta, A.; Vedaldi, A.; and Zisserman, A. 2016. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2315–2324.

Gupta, A.; Vedaldi, A.; and Zisserman, A. 2018. Learning to read by spelling: Towards unsupervised text recognition. In *Proceedings of the 11th Indian Conference on Computer Vision, Graphics and Image Processing*, 1–10.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2021. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Howard, R.; Brammer, M.; David, A.; Woodruff, P.; Williams, S.; et al. 1998. The anatomy of conscious vision: an fMRI study of visual hallucinations. *Nature neuroscience*, 1(8): 738–742.

Hradiš, M.; Kotera, J.; Zemcık, P.; and Šroubek, F. 2015. Convolutional neural networks for direct text deblurring. In *Proceedings of BMVC*, volume 10.

I. Pratikakis, K. N., B. Gatos. 2011. ICDAR 2011 document image binarization contest (DIBCO 2011). In *2011 International Conference on Document Analysis and Recognition*, 1506–1510.

Jemni, S. K.; Souibgui, M. A.; Kessentini, Y.; and Fornés, A. 2022. Enhance to read better: A Multi-Task Adversarial Network for Handwritten Document Image Enhancement. *Pattern Recognition*, 123: 108370.

Kang, L.; Riba, P.; Rusiñol, M.; Fornés, A.; and Villegas, M. 2020a. Pay attention to what you read: Non-recurrent handwritten text-line recognition. *arXiv preprint arXiv:2005.13044*.

Kang, L.; Rusinol, M.; Fornés, A.; Riba, P.; and Villegas, M. 2020b. Unsupervised writer adaptation for synthetic-to-real handwritten word recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3502–3511.

Kang, S.; Iwana, B. K.; and Uchida, S. 2021. Complex image processing with less data—Document image binarization by integrating multiple pre-trained U-Net modules. *Pattern Recognition*, 109: 107577.

Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; i Bigorda, L. G.; Mestre, S. R.; Mas, J.; Mota, D. F.; Almazan, J. A.; and De Las Heras, L. P. 2013. ICDAR 2013 robust reading competition. In *2013 12th International Conference on Document Analysis and Recognition*, 1484–1493. IEEE.

Kleber, F.; Fiel, S.; Diem, M.; and Sablatnig, R. 2013. Cvl-database: An off-line database for writer retrieval, writer identification and word spotting. In *2013 12th international conference on document analysis and recognition*, 560–564. IEEE.

Kolesnikov, A.; Zhai, X.; and Beyer, L. 2019. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1920–1929.

Liang, D.; Li, L.; Wei, M.; Yang, S.; Zhang, L.; Yang, W.; Du, Y.; and Zhou, H. 2022. Semantically contrastive learning for low-light image enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1555–1563.

Litman, R.; Anschel, O.; Tsiper, S.; Litman, R.; Mazor, S.; and Manmatha, R. 2020. Scatter: selective context attentional scene text recognizer. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11962–11972.

Liu, H.; Wang, B.; Bao, Z.; Xue, M.; Kang, S.; Jiang, D.; Liu, Y.; and Ren, B. 2022. Perceiving Stroke-Semantic Context: Hierarchical Contrastive Learning for Robust Scene Text Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI.

Long, S.; He, X.; and Yao, C. 2021. Scene text detection and recognition: The deep learning era. *International Journal of Computer Vision*, 129(1): 161–184.

Lu, H.; Kot, A. C.; and Shi, Y. Q. 2004. Distance-reciprocal distortion measure for binary document images. *IEEE Signal Processing Letters*, 11(2): 228–231.

Marti, U.-V.; and Bunke, H. 2002. The IAM-database: an English sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1): 39–46.

Memon, J.; Sami, M.; Khan, R. A.; and Uddin, M. 2020. Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR). *IEEE Access*, 8: 142642–142668.

Mishra, A.; Alahari, K.; and Jawahar, C. 2012. Scene text recognition using higher order language priors. In *BMVC-British machine vision conference*. BMVA.

Noroozi, M.; and Favaro, P. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, 69–84. Springer.

Ntirogiannis, K.; Gatos, B.; and Pratikakis, I. 2012. Performance evaluation methodology for historical document image binarization. *IEEE Transactions on Image Processing*, 22(2): 595–609.

Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; and Efros, A. A. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2536–2544.

Pratikakis, I.; Gatos, B.; and Ntirogiannis, K. 2012. ICFHR 2012 competition on handwritten document image binarization (H-DIBCO 2012). In *2012 international conference on frontiers in handwriting recognition*, 817–822. IEEE.

Pratikakis, I.; Zagori, K.; Kaddas, P.; and Gatos, B. 2018. ICFHR 2018 Competition on Handwritten Document Image Binarization (H-DIBCO 2018). In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 489–493.

Pratikakis, I.; Zagoris, K.; Barlas, G.; and Gatos, B. 2016. ICFHR2016 handwritten document image binarization contest (H-DIBCO 2016). In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 619–623. IEEE.

Pratikakis, I.; Zagoris, K.; Barlas, G.; and Gatos, B. 2017. ICDAR2017 competition on document image binarization (DIBCO 2017). In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, 1395–1403. IEEE.

Sauvola, J.; and Pietikäinen, M. 2000. Adaptive document image binarization. *Pattern recognition*, 33(2): 225–236.

Shi, B.; Bai, X.; and Yao, C. 2016. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11): 2298–2304.

Shi, B.; Wang, X.; Lyu, P.; Yao, C.; and Bai, X. 2016. Robust scene text recognition with automatic rectification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4168–4176.

Smith, R. 2007. An overview of the Tesseract OCR engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, 629–633. IEEE.

Sonkusare, M.; and Sahu, N. 2016. A survey on handwritten character recognition (HCR) techniques for English alphabets. *Advances in Vision Computing: An International Journal (AVC)*, 3(1).

Souibgui, M. A.; Biswas, S.; Jemni, S. K.; Kessentini, Y.; Fornés, A.; Lladós, J.; and Pal, U. 2022. DocEnTr: An End-to-End Document Image Enhancement Transformer. *arXiv preprint arXiv:2201.10252*.

Souibgui, M. A.; Fornés, A.; Kessentini, Y.; and Megyesi, B. 2021. Few Shots Is All You Need: A Progressive Few Shot Learning Approach for Low Resource Handwriting Recognition. *arXiv preprint arXiv:2107.10064*.

Souibgui, M. A.; and Kessentini, Y. 2020. DE-GAN: a conditional generative adversarial network for document enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Souibgui, M. A.; Kessentini, Y.; and Fornés, A. 2020. A conditional GAN based approach for distorted camera captured documents recovery. In *Mediterranean Conference on Pattern Recognition and Artificial Intelligence*, 215–228. Springer.

Sueiras, J.; Ruiz, V.; Sanchez, A.; and Velez, J. F. 2018. Offline continuous handwriting recognition using sequence to sequence neural networks. *Neurocomputing*, 289: 119–128.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vincent, P.; Larochelle, H.; Bengio, Y.; and Manzagol, P.-A. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, 1096–1103.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8798–8807.

Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; and Deny, S. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, 12310–12320. PMLR.

Zhang, R.; Isola, P.; and Efros, A. A. 2016. Colorful image colorization. In *European conference on computer vision*, 649–666. Springer.

Zhang, X.; Zhu, B.; Yao, X.; Sun, Q.; Li, R.; and Yu, B. 2022. Context-based Contrastive Learning for Scene Text Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI.

Zhang, Y.; Nie, S.; Liu, W.; Xu, X.; Zhang, D.; and Shen, H. T. 2019. Sequence-to-sequence domain adaptation network for robust text image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2740–2749.

Zhao, J.; Shi, C.; Jia, F.; Wang, Y.; and Xiao, B. 2019. Document image binarization with cascaded generators of conditional generative adversarial networks. *Pattern Recognition*, 96: 106968.

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

A Overview

The main intuition behind the Text-DIAE framework has been the degradation invariant features learnt during pre-training without any manually annotated samples, which is fine-tuned for key downstream application tasks: Text-Recognition and Document Image Enhancement. In this supplementary material, we discuss and highlight some essential insights of Text-DIAE under the two sections B and C.

B Text Recognition

B.1 Transformations

We begin showing the augmentations employed to the input images used at the pre-training stage. We carefully select transformations suited for text in order to not disrupt the sequential information of characters found at the word level.

Results of transformations employed can be seen in Figure 6. The transformations used are: addition of Gaussian noise, shear, minor rotation, scaling and cropping upon some detected points given by a mask in order to not alter the characters in a significant manner. Code regarding the data augmentation pipeline employed will be made public at UponAcceptance.

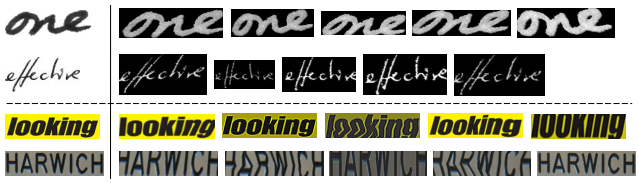


Figure 6: **Transformations during training.** Image augmentations are carefully selected in order to not disrupt the sequentially nature of characters contained in a word.

B.2 Datasets

In this work, we make use of the following public datasets to train models in handwriting and scene text recognition.

- **IAM** (Marti and Bunke 2002): It is a handwritten English text dataset. It contains 657 different writers, and it is partitioned into writer independent training, validation and test. Addtionally, this dataset is comprised of 74,805 fully segmented words.
- **CVL** (Kleber et al. 2013): Dataset of handwritten text in English language. It was written 310 different writers, partitioned into writer independent training and test sets. In total, 27 writers wrote 7 texts and the remaining 283 writers wrote 5 texts.
- **IIIT5K-words** (IIIT5K) (Mishra, Alahari, and Jawahar 2012): This is a dataset made of 2000 training and 3000 testing cropped scene text images from the Internet. Often employed in recognition and scene text retrieval tasks.
- **ICDAR-2013** (IC13) (Karatzas et al. 2013): It is a dataset that contains 848 training and 1015 testing cropped scene text in focused environments, where the text was relevant. Often employed in text detection and recognition.
- **Modified-SynthText**: This dataset is a modified version of (Gupta, Vedaldi, and Zisserman 2016). It was employed by (Gómez et al. 2018) to train a scene-text retrieval model. It contains 4M cropped scene text images which were generated synthetically. This dataset is smaller than the original one and it avoids random flipping of text images and rotated text that can break the sequential order of characters.

The pre-training datasets are domain-dependant, in handwritten text we used IAM and CVL. In scene text we used the Modified-SynthText as well as the training sets from ICDAR13 (Karatzas et al. 2013) and IIIT5K (Mishra, Alahari, and Jawahar 2012).

B.3 Evaluation metrics

Character Error Rate (CER). CER is used to compare the similarity of two strings at character level. Given two stings s_1 and s_2, where s_2 is the reference (ground truth):

$$CER(s_1, s_2) = \frac{S + D + I}{N} \tag{5}$$

where S is the number of substitutions, D is the number of deletions and I is the number of insertions that are applied in s_1 to obtain $s_1 = s_2$. Noting that, N is the number of characters in s_2.

Edit Distance at 1 (ED1). It represents the accuracy of a prediction up to a edit distance of 1. This implies that a prediction is correct compared to the ground truth if only *one* substitution, deletion or insertion is required for both strings to match.

B.4 Implementation details

In order to provide the reader with reproducibility of the presented work, the implementation details of the models showcased for Handwritten and Scene Text Recognition are shown and summarized in Table 8.

During pre-training we deploy an encoder with 6 layers and 8 attention heads to encode the input, with a dimension of 768. We used this same number of layers and attention heads for the decoder, with a dimension of 512 in text recognition and 768 for document enhancement. At masking, each input image (with size $64 \times 256 \times 3$ for text recognition and $256 \times 256 \times 3$ for document enhancement) is divided into a set of patches with size $8 \times 8 \times 3$. Similarly as (He et al. 2021), we employ random masking of 75% of the patches. To add blur, we add average blur with random kernel sizes between 1 and 15. In order to add background noise, we add weighted contrasting backgrounds from different text documents. The pre-training was done for 2 epochs in scene text and 100 epochs for the handwritten domain.

In the fine tuning stage, we use the same encoder (pre-trained) with a different decoder of 6 layers, 8 attention heads and dimension of 768.

B.5 Qualitative Insights

In Figure 7 we show qualitative samples of the original, masked and reconstructed images at pre-training. The images correspond to the validation set and the model have not

Table 8: **Implementation Details.** Implementations details of Handwritten and Scene Text Recognition. The acronyms STR and HTR stands for Scene Text Recognition and Handwritten Text Recognition respectively.

ConFigure	Pre-Training	Fine tuning	Scratch
optimizer	AdamW	Adam	Adam
learning rate	1.5 e^{-4}	1 e^{-4}	1.5 e^{-5}
weight decay	0.05	0.05	0.05
optimizer momentum	β_1, β_2=0.9, 0.95	β_1, β_2=0.9, 0.95	β_1, β_2=0.9, 0.95
batch size	64 (HTR) / 192 (STR)	64 (HTR) / 256 STR	64 (HTR) / 256 STR
learning rate schedule	cosine decay	cosine decay	cosine decay
warmup epochs	3	3	10
training epochs	100 (HTR) / 2 (STR)	600 (HTR) / 10 (STR)	600 (HTR) / 10 (STR)

Figure 7: **Reconstruction of Handwritten Text.** Qualitative samples of reconstructed input images. The model may have seen the word, but it has not seen the images at training. The reconstruction is performed after employing masking as pretext task during pre-training.

seen them. In Figure 7, we focus on the pre-text task (masking - 75%) that yields the biggest improvement on the downstream task. We note that at pre-training the model learns a distribution of characters (English) that helps to solve the masking task. It is worth noting the capability of keeping similar font types and applying a padding in short words. In harder cases, such as in the words "stucked", "saying" and "having" the model still reconstructs blurry or wrong characters. However, we have to note the complexity of this task, which is very demanding even for humans.

Similarly, we show qualitative samples of reconstructing images at pre-training by masking an input scene-text cropped word in Figure 8. The model has not seen this images before. The model learns to properly reconstruct input images, while properly keeping the background color and specific font style. Some failure cases are shown while reconstructing "pea" and "earthsellers".

We show in Figure 9 the recognized text in the test sets of IAM and CVL datasets. Despite the difficulty of some samples, the model achieves state-of-the-art in IAM. Failure cases depict problematic handwritten styles to be recognized, e.g. "tonight" and "Christmas". Some characters are mistakenly predicted, thus causing errors in the next sequence of characters, such in "irresolute" and "wiener". Some errors come from difficult cropped samples such as "strikes", in which the sample starts with a strikethrough word.

C Document Image Enhancement

C.1 Datasets

- **DIBCO and H-DIBCO datasets**: These datasets were introduced in the Document Image Binarization challenges that took place since 2009. In this work we have mainly chosen two DIBCO(2011 (I. Pratikakis 2011) and 2017 (Pratikakis et al. 2017)) and two H-DIBCO(2012 (Pratikakis, Gatos, and Ntirogiannis 2012) and 2018 (Pratikakis et al. 2018)) benchmarks for final validation on the document image binarization task. All the image samples are mainly historical documents that has undergone several distortions like bleed, show-through, smears, fading and so on. The DIBCO 2011 and 2017 contained 16 and 20 samples respectively. While the H-DIBCO 2012 and 2018 contained 14 and 10 instances respectively.
- **Blurry Document Images**: The training dataset con-

Figure 8: **Reconstruction of Scene Text.** Qualitative samples of reconstructed input images. The model may have seen the word, but it has not seen the images at training. The reconstruction is performed after employing masking as pretext task during pre-training.

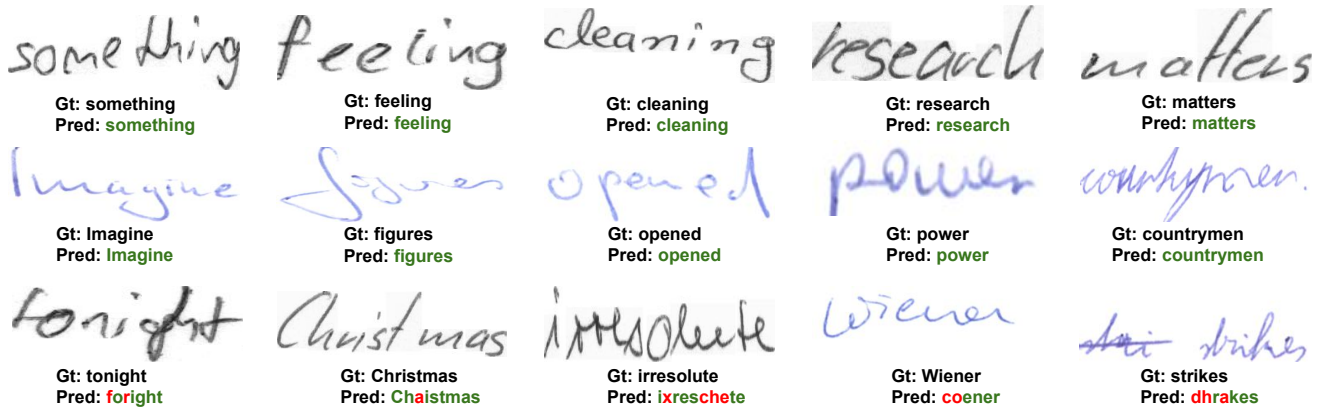


Figure 9: **Handwritten Text Recognition.** Qualitative samples of attaching a decoder and fine-tuning an encoder to perform recognition of IAM (black font) and CVL (blue font). Correctly predicted samples are displayed in green, while mistakes are shown in red. Failure cases are depicted on the last row.

tains 4000 training images and 932 images for validation of resolution 300x300 patches. This data was used in (Souibgui and Kessentini 2020) and it is originally a subset from the dataset proposed in (Hradiš et al. 2015). Every instance is extracted from a different document image and each blur kernel used is exclusive.

For pre-training we used 203,576 unlabeled document image samples. These samples were taken from the historical document DIBCO benchmarks (2009, 2010, 2013, 2014, 2015, 2016) (Pratikakis et al. 2016), Publaynet (Zhong, Tang, and Yepes 2019), Palm-leaf (Burie et al. 2016) and IAM Handwriting database.

C.2 Evaluation metrics

Peak signal-to-noise ratio (PSNR). PSNR helps to establish a pixel-wise validation and measures the effectiveness of document image enhancement approaches in terms of visual quality. It computes the ratio between the maximum possible value of the signal(image) to the amount of noise (distortion) that affects the quality. The higher the value, the more similar are the two images which in turn assures that the reconstructed image (binarized or deblurred) has a better quality. MAX is the maximum possible pixel value of the image (eg. MAX is 255 for pixels represented as 8 bits per sample) . Given two $M * N$ images, they can be fomulated as shown in eqn. 6.

$$MSE = \frac{\sum_{M,N} [I_1(m,n) - I_2(m,n)]^2}{M * N}$$
$$PSNR = 10 \log_{10} \left(\frac{MAX^2}{MSE} \right) \quad (6)$$

F-Measure(FM). The F-measure gives the harmonic mean between the precision (P) and recall (R) scores. Precision determines the number of positive predictions, while recall measures the ability to find the positive predictions in a binary classifier. Here, in binarization problem, FM com-

putes the accurate prediction of the white background and the black foreground (text) between a binarized sample and the ground-truth(GT) sample. FM is formulated as shown in eqn. 7.

$$FM = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \tag{7}$$

Pseudo F-Measure (F_{ps}). The Pseudo F-Measure was first introduced for document image binarization task in (Ntirogiannis, Gatos, and Pratikakis 2012) as it proposed a normalization procedure on the weights of the GT image(both background and foreground) for a lower penalization.

Distance Reciprocal Distortion (DRD). DRD was mainly used to measure visual distortion for all pixels in binary document images as proposed in (Lu, Kot, and Shi 2004). It is computed at first kth flipped pixel using a normalized weight matrix (5x5) W_{norm} defined in the same paper (Lu, Kot, and Shi 2004). The formulation is done as shown in eqn. 8. To calculate the overall DRD score between the binarized resultant image $B_k(x, y)$ and the GT image $GT_k(i, j)$, it is basically to sum up the DRD for N number of flipped pixels starting from $k = 1$ and NB_{nu} refers to the number of non-uniform blocks as formulated in eqn. 9.

$$DRD_k = \sum_{i=-2}^{2} \sum_{j=-2}^{2} |GT_k(i, j) - B_k(x, y)| \times W_{norm}(i, j) \tag{8}$$

$$DRD = \frac{\sum_{k=1}^{N} DRD_k}{NB_{nu}} \tag{9}$$

C.3 Implementation details

The implementation details of the models showcased for Document Image Enhancement tasks (binarization and deblurring) are shown and summarized in Table 9. The Xavier-Glorot uniform (Glorot and Bengio 2010) was used to initialize the transformer blocks, as implemented in the original ViT (Dosovitskiy et al. 2021). The ViT-Base model variant was used as the main model baseline with the patch size 8 and Random resized Crop was used as the default applied transformation for all the images during pre-training. The pre-training was done for 100 epochs.

C.4 Qualitative Insights

The visual appearance of a historical document has been negatively affected with time by different forms of degradation. Recovering the document with its foreground text and information is the key challenge in document image enhancement. In this section, we will discuss some qualitative highlights of our proposed Text-DIAE model applied to document image binarization. We shall also cover some analysis on the task of document deblurring mainly on camera-captured document images affected by motion, out-of-focus blurs etc.

Table 9: **Training settings:** Text-DIAE model configurations and hyperparameters adapted for Document Image Enhancement tasks.

ConFigure	Pre-Training	Fine tuning
optimizer	AdamW	AdamW
learning rate	1.5e-4	1.5e-4
weight decay	0.05	0.05
optimizer momentum	(0.9,0.95)	(0.9,0.99)
batch size	64	64
learning rate schedule	cosine	cosine
warmup epochs	3	15
training epochs	50	100

Text-DIAE recovers show-through. The show-through problem appears when ink impressions from one side of the paper start appearing on the other side, making the document almost illegible. As shown in the results of examples (2nd, 3rd, 4th, 5th and 6th) of Figure 11 Text-DIAE binarized results get rid of the show-through successfully to make the document easily readable. An other qualitative comparison with other approaches (refer to Figure 12) reveals the superiority of our approach than the related work in this task.

Text-DIAE recovers smears or stains. As shown in the 1st example of Figure 11 documents also suffer from stains or smears and need to be recovered for proper readability. Text-DIAE not only successfully recovers the text (foreground) but also smoothens its boundary pixels.

Text-DIAE recovers faint characters or weak text. Every individual glyph (character) inside a document can either appear faded due to the ink quality or paint as they start shrinking with time. Even adaptive handcrafted binarization approaches fail to recover and extract the text accurately. We illustrate in Figure 10 an example of thin parchment of text and run the Tesseract-OCR engine (with their already in-built adaptive binarization tool) for validating the Text-DIAE model. The predicted OCR output from the original input image is compared with the binarized results from Text-DIAE, DocEnTr and the ground-truth image. The improvement in CER when compared with the original input and DocEnTr proves that Text-DIAE works substantially well to recover weak text for an OCR to read. Also, in the 8th example of Figure 11, Text-DIAE performs quite good for thin characters under a faded background.

Text-DIAE recovers contrast variations. Most of the times in historical degraded documents, they suffer from huge differences between high/low pixels due to occlusion, illumination and other noisy factors. In the examples shown in Figure 11 the binarized results have been obtained under several different contrasting background variations. This can be attributed to the feature representations learnt during pre-training stage by Text-DIAE by applying background noise transformation.

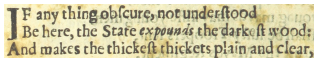
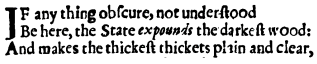
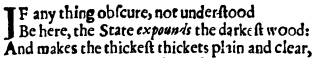
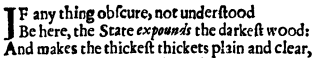
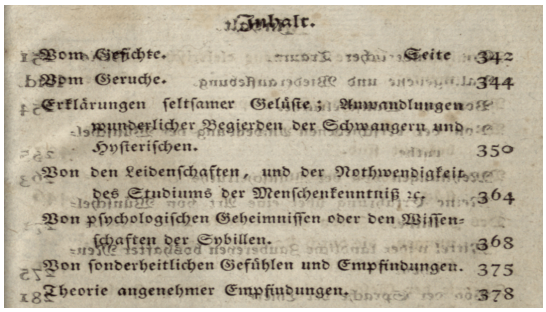
Original Input	DocEnTr (Souibgui et al. 2022)	**Ours**	Ground Truth
			
OCR output: "Janythingobfcure, notunderftood,,., ... ,Be here, the State expounds the darkeft wood: And makes the thickeft thick-ets p!ain-and clear,"	**OCR output:** "Iany thing obfcure,notunder ftood Be here, the State expounds the darkeft wood: Aod makesche thickeft thickets p!tinand clear,"	**OCR output:** "Ijany thing obfcure,not underftood Be here, the State expounds the darkeft wood: And makesche thickeft thick-ets pltinand clear,"	**OCR output:** "Fany thing obfcure,not underftood Be here, the State expounds the darkeft wood: And makesche thickeft thick-ets p!ainand clear,"
CER: 17.97	CER: 13.28	CER: **11.50**	CER: 10.94

Figure 10: **Qualitative analysis of binarized samples for OCR.** The document image on the left refers to the originally captured image, followed by the DocEnTr (Souibgui et al. 2022) binarized result, and the Text-DIAE binarized result, and the ground-truth image in the last column. The correctly predicted OCR output is shown in "Green" font while the inaccurate ones are depicted in "Red" and recognition performance in terms of CER.

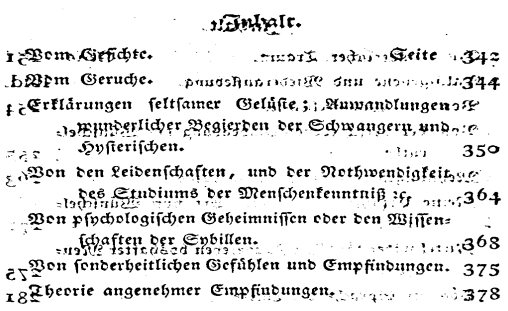
Text-DIAE excels in deblurring task. When it comes to camera-captured documents, it often suffers from different kinds of blurring. Motion blur artefacts occur due to the relative speed between the object and the camera, where a sudden camera movement leads to a degradation in the captured image. The out-of-focus blur mainly occur when light fails to converge in the during camera capture. In this work, Text-DIAE was mainly applied to the benchmark data proposed in (Hradiš et al. 2015) to see how the model performs in document deblurring. The results shown in Figure 13 exhibit how the deblurred results from Text-DIAE achieves almost ground-truth level recovery from both motion-blur and out-of-focus blurring effects. The enhancement of the blurred images also help to capture the text from the deblurred sample, an illustration of which has been already presented in the main section of the paper. Also, the deblurring learning objective which is applied during the pre-training stage contributes significantly to improve the performance on the blurred document samples.

Robustness Evaluation of Text-DIAE. We use the same model pre-trained on the DIBCO datasets (2009, 2010, 2013, 2014, 2015, 2016), Palm Leaf images (Burie et al. 2016) and IAM (Marti and Bunke 2002) handwriting database and use it on fine-tuning for two different document image enhancement tasks: binarization and deblurring, without any special change in settings. This proves the robustness of Text-DIAE as it learns universal feature representations from different kinds of degradation (background noise, blur and masking) beneficial for multiple downstream tasks.

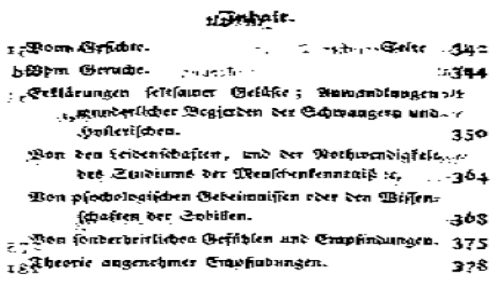
Figure 11: **Qualitative analysis of Text-DIAE in binarization task.** Images in columns are: Left: Original image, Middle: Binarized image using our proposed method., Right: Ground Truth image



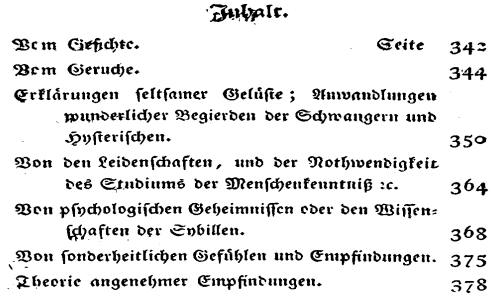
Original



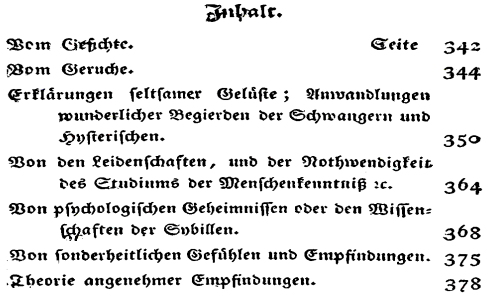
(Sauvola and Pietikäinen 2000)



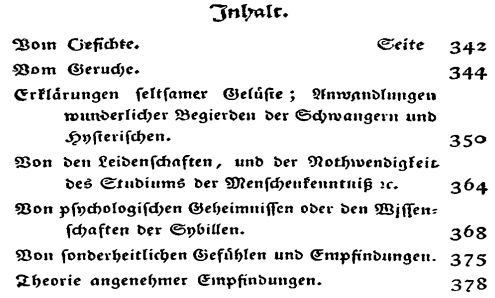
(Kang, Iwana, and Uchida 2021)



(Souibgui et al. 2022)



Ours



Ground Truth

Figure 12: **Qualitative results of binarized samples:** We show the results of Text-DIAE on the document image binarization task. Given a degraded input example from DIBCO 2017, our model performs significantly better qualitatively compared to previous approaches.

Figure 13: **Qualitative analysis of Text-DIAE in deblurring task.** Images in columns are: Left: Original image, Middle: Deblurred image output from Text-DIAE, Right: Ground Truth image