DocNLC: A Document Image Enhancement Framework with Normalized and Latent Contrastive Representation for Multiple Degradations

Ruilu Wang, Yang Xue*, Lianwen Jin

School of Electronic and Information Engineering, South China University of Technology
ruiluwang2rylonw@gmail.com, {yxue, eelwjin}@scut.edu.cn

Abstract

Document Image Enhancement (DIE) remains challenging due to the prevalence of multiple degradations in document images captured by cameras. In this paper, we respond an interesting question: can the performance of pre-trained models and downstream DIE models be improved if they are bootstrapped using different degradation types of the same semantic samples and their high-dimensional features with ambiguous inter-class distance? To this end, we propose an effective contrastive learning paradigm for DIE — a **Doc**ument image enhancement framework with **N**ormalization and **L**atent **C**ontrast (DocNLC). While existing DIE methods focus on eliminating one type of degradation, DocNLC considers the relationship between different types of degradation while utilizing both direct and latent contrasts to constrain content consistency, thus achieving a unified treatment of multiple types of degradation. Specifically, we devise a latent contrastive learning module to enforce explicit decorrelation of the normalized representations of different degradation types and to minimize the redundancy between them. Comprehensive experiments show that our method outperforms state-of-the-art DIE models in both pre-training and fine-tuning stages on four publicly available independent datasets. In addition, we discuss the potential benefits of DocNLC for downstream tasks. Our code is released at https://github.com/RylonW/DocNLC.

Introduction

Document preservation can be hampered by a variety of degradations such as watermarks, blurring, noise, backgrounds, ink bleed, and uneven ambient light (Lin, Chen, and Chuang 2020; Anvari and Athitsos 2021). As a result, a number of deep learning-based correction methods have been proposed for specific types of degradation, including document watermark removal (Souibgui and Kessentini 2020) , blurring removal (Ljubenović and Figueiredo 2019; Tran et al. 2021) , noise removal (Gangeh et al. 2021) , and shadow removal (Lin, Chen, and Chuang 2020) . However, most of them focus on a specific type of degradation and are therefore less applicable in practice to other types of degradation.

*Corresponding author: Yang Xue

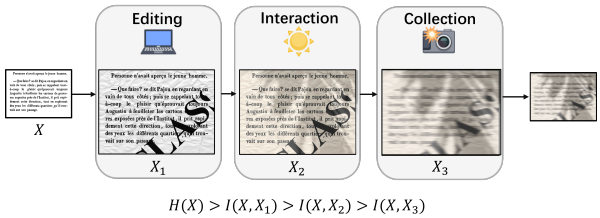


Figure 1: The depiction of the document image degradation process provides a valuable guideline for our contrastive learning design. According to data processing inequality theory (Cover 1999), the amount of mutual information $I(x, \cdot)$ between a document image and the original sample decreases over a series of editing, interaction, and collection processes. Therefore, it is crucial not to simply mix the datasets, but to focus on exploring the relationships between them.

The process of document image degradation can be divided into three distinct phases: editing, interaction and collection. In the editing phase, the background of the document varies, depending on the material used to write or print it. Also, the watermarks could be added. In the interaction phase, the document interacts with the environment in which it is placed. The light in the physical environment and the moisture in the air will continue to interact with the document entity, causing shadows, wrinkles and corrosion (Ma et al. 2018; Das et al. 2019; Li et al. 2019; Xie et al. 2020; Feng et al. 2021). In addition, human behaviour can exacerbate the degradation of the document image, such as stamping images or making notes on the original document. Finally, during the collection phase, motion blurring in different directions can occur due to instability in the shooting equipment. Thermal noise in the camera electronics or multiple compressions and decompressions of the image during storage and transmission can also introduce noise, further degrading the quality of document images. The degradation of document images is therefore a multi-stage process. This complex degradation process creates difficulties in reading documents and challenges in recovering document images. The simplest way to solve this problem is to train specific networks for each type of degra-

dation (He and Schomaker 2019; Ljubenović and Figueiredo 2019; Souibgui and Kessentini 2020; Lin, Chen, and Chuang 2020). Among the above methods, DE-GAN (Souibgui and Kessentini 2020) is the most representative and complete work, where separate GAN models with the same structure are trained for binarization, watermarks and blurring, but this leads to a significant increase in training time and model parameters. The second solution is to train the network with a mixture of data from different types of degradation to improve its enhancement capability. However, the representation of different types of degradation varies, as does the distance between each type of degradation and the original sample. If the mixed data is used directly for training, the relationship between the different types of degradation is ignored, resulting in a poor model.

Therefore, inspired by exposure normalization correction (Huang et al. 2022) and contrastive learning for image enhancement (Liang et al. 2022), we propose a framework to handle multiple types of document degradation in a unified way. There are numerous existing degraded document datasets, but most of them contain only a single degradation type. To accommodate our proposed framework, we merge several publicly available datasets and artificially fill in the missing types so that each original document sample has the corresponding five degradation types: backgrounds, watermarks, shadow, noise and blur. Following the basic principle of normalization followed by compensation, our degradation normalization model starts with a rough alignment of different document degradation features, followed by the calculation of the normalization loss. However, normalization inevitably loses some features during image reconstruction, so the compensation part integrates the features pre- and post-normalization to ensure the integrity of the information. In addition, since the pre-training is performed on synthetic datasets, we fine-tune the model on real datasets to improve the performance of the model on real degraded document images. This strategy reduces the difficulty of repairing different classes of document degradation and facilitates the formation of a network that can handle multiple degradations in a balanced manner.

The main contributions of this work are summarized as follows:

1. Unlike most existing DIE methods that train a specific network for each type of degradation, DocNLC proposed in this paper is the first unified document image enhancement framework that deals with multiple types of degradation. Through direct and latent contrastive learning, DocNLC can extract the common features of different degradation types and reduce their redundancy, so that document images of different degradation types can be uniformly restored to clean images.

2. Existing contrastive learning frameworks increase the inter-class distance, but the distance between different types of degraded image features is ambiguous, leading to class separation difficulties and low-quality image restoration. To address this problem, we propose a different contrast learning paradigm that enhances explicit deconvolution between different types of degraded representations. Using inter-class relations, different types of degraded features are mapped into degradation-independent representations.

3. Comprehensive experimental results on four publicly available independent datasets show that DocNLC outperforms the state-of-the-art and demonstrate the generality and effectiveness of the framework proposed in this paper.

Related Work

Document Imgae Enhancement

Initially, researchers considerd Document Image Enhancement (DIE) as a binarization task and used sliding windows and threshold segmentation strategies (Otsu 1979; Sauvola and Pietikäinen 2000) to solve this problem. However, the introduction of U-Net (Ronneberger, Fischer, and Brox 2015) sparked a huge research interest in this area. DeepOtsu (He and Schomaker 2019) was the first paper to combine thresholding with U-Net(Ronneberger, Fischer, and Brox 2015). Subsequently, Kang et al. (Kang, Iwana, and Uchida 2021) proposed a stack of pre-trained U-Net modules from a data-driven perspective. In addition, generative adversarial network-based approaches (Zhao et al. 2019; Souibgui and Kessentini 2020) have been that treat DIE as an image generation task. These approaches employ a generator to produce an enhanced version of the document image, while a discriminator is used to evaluate the quality of binarization. More recently, attempts have been made to relate DIE to other tasks such as optical character recognition, which requires access to an effective upstream representations. Consequently, Transformer-based techniques, such as DocEnTr (Souibgui et al. 2022) and Text-DIAE (Souibgui et al. 2023), have been proposed to recover the encoder representation. However, while these methods appear to address a range of degradation types, none of them have been thoroughly tested for cross-dataset performance. An exception is BEDSR-Net (Lin, Chen, and Chuang 2020), which uses U-Net and discriminators to specifically correct for shadows in document images. Finally, Kligler et al. (Kligler, Katz, and Tal 2018) presented a new interpretation of DIE, defining it as a 3D point visibility detection task. However, this approach also fails to take into account the practical application scenarios of DIE, which involves multiple types of degradation.

Contrastive Learning

In recent years, contrastive learning (CL) has become a prominent approach for learning invariant representations by focusing on feature differences. The prevailing methods in this field mainly rely on the use of negative samples(Oord, Li, and Vinyals 2018; He et al. 2020; Chen et al. 2020). However, these negative sample-based methods have strict requirements on the number of negative samples per batch, resulting in an increased storage burden. To overcome this limitation, several alternative contrastive methods have been proposed. Instead of relying on weight sharing between branches, output quantization, stop gradient, memory banks, and other complex techniques, these methods take a different perspective, using explicit decorrelation as a means of learning representations(Grill et al. 2020; Zbontar et al. 2021; Zhang et al. 2021; Bardes, Ponce, and LeCun 2022).

By simplifying the CL strategy, these methods eliminate the need for cumbersome techniques and make the learning process more streamlined and efficient. In addition, certain approaches address the challenge from a more novel perspective, such as SFA (Yifei et al. 2023), which rebalances the partial contribution on the feature spectrum to improve the quality of representations, thus providing a unique solution to this problem.

Methodology

Motivation and Overview

As shown in Figure 1, correcting different levels of degradation has different visual characteristics, making it very difficult for the network to distinguish between foreground and background for various degradation types. In addition, training on mixed datasets not only increases the training time, but also promotes superficial memory, leading to unbalanced network performance across different datasets and degradation types. Therefore, this paper proposes to deal with multiple degradation types under a unified framework (DocNLC) that exploits the relationship between different degradations and learns content-invariant degradation representations using direct and latent contrastive learning methods.

Figure 2 shows the architecture of DocNLC. It consists of three main components: the degradation normalization and compensation module (DNC), the restoration module, and the latent contrastive module. When a degraded document is input, the DNC module is applied first. The compensated output is sent to the subsequent restoration module, while the normalized output is passed to the latent contrastive module. Importantly, the latent contrastive module is only used during training and does not introduce any additional parameters into the final network.

In terms of task modules, we use two widely-used networks: the restoration module is similar to the UNet architecture and is responsible for pixel-level correction and remapping. On the other hand, the latent contrastive module draws inspiration from the Barlow Twins (Zbontar et al. 2021) framework. We adopt two separate streams to address the dual aspects of the latent contrastive module - contrastive learning and content preservation.

Degradation Normalization and Compensation

Figure 3 shows the architecture of the DNC module. The normalization part aims to map various degraded document image features into a feature-invariant space, while the compensation part integrating the unprocessed features to compensate for the lack of discriminative information in the image caused by normalization.

Normalization Module The features are first roughly aligned using instance normalization. Assuming that the input feature x , we perform instance normalization(Huang and Belongie 2017) using Eq. (1):

$$\mathcal{N} = IN(x) = \gamma \frac{x - \mu(x)}{\sigma(x)} + \beta \tag{1}$$

where $\mu()$ and $\sigma()$ denote the mean and standard deviation computed across spatial dimensions for each channel and each sample, γ and β are learnable parameters.

By integrating instance normalization within the model architecture, the input feature statistics can be efficiently normalized. Different degradations are aligned using instance normalization, thus reducing their representation discrepancies. The output of the normalization part is fed to the projector head as shown in Fig.2. The normalization part effectively close the distance between the degradation features, so that the gap between the new feature maps processed by the projector is not too large to be optimized.

Compensation Module Normalization process inevitably removes information, resulting in insufficient information for image reconstruction. To overcome this drawback, we add a compensation part to integrate the initial features not processed by the normalization component to ensure the completeness of the information that can be sent to the subsequent restoration module. Specifically, we implement the compensation in both spatial and channel dimensions, which helps to guide the integration of missing information in the initial features. This idea was inspired by SENet (Hu, Shen, and Sun 2018).

Contrastive Degradation Representation

DocNLC addresses a general document restoration problem, rather than a specific type of degradation problem. This requires a large amount of diverse data to be used during training. However, naively feeding the data to the network without exploiting the inter-class relationships is suboptimal. We propose a contrastive learning framework that links multiple degradation types and enhances the network's generalization ability. Contrary to existing CL frameworks(e.g. Sim-CLR) that emphasize negative samples and aim to create a linearly separable space for different image categories, we argue that this is not suitable for our task. Since real degraded document images contain multiple degradation types, enforcing large distance between categories (different degradation types) may degrade the restoration quality. What's more, unlike existing symmetric CL frameworks, DocNLC assigns different importance to the primary and other modalities, resulting in an asymmetric contrastive form. This is shown in the bottom left of Figure 2. To this end, we adopt a different CL paradigm that enforces explicit decorrelation of the representations. Before explaining our contrastive algorithm, we define the input of the primary modality as x^m and the inputs of other modalities as x^{m1}, x^{m2}, etc. Similarly, their representation for latent contrast are defined as r^m, r^{m1}, r^{m2}, etc. The complete pytorch-style pseudo-code for our constrastive learning strategy is shown in Algorithm 1.

Latent Contrastive Constraint We aim to project the normalized features (Ioffe and Szegedy 2015) into another dimension and compute the correlation matrix with maximum invariance and minimum redundancy. Unlike Barlow Twins, we do not need to increase the feature dimensions after normalization, but reduce them to speed up the computation. Our method reduces the feature dimensions in both the

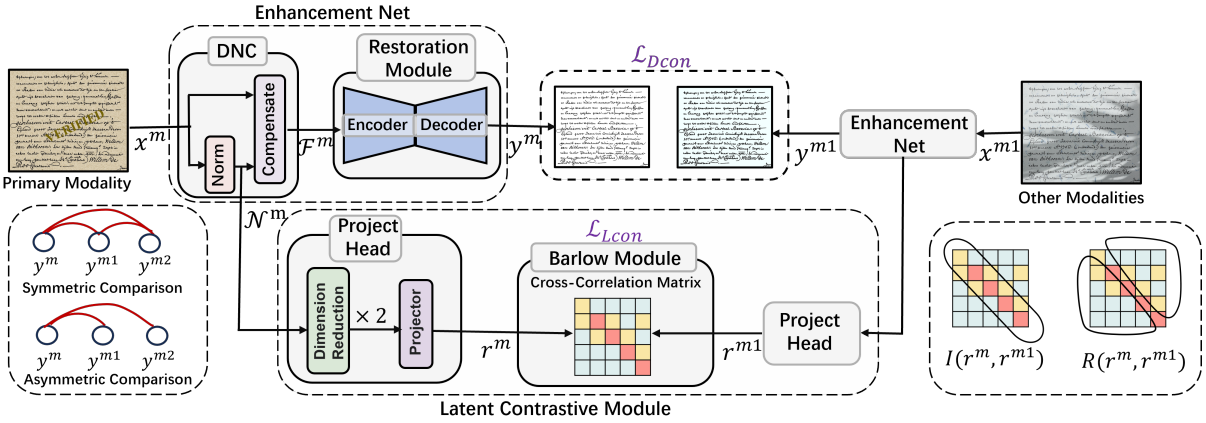


Figure 2: The overall architecture of DocNLC. In the training phase, the latent contrastive module plays a vital role in parameter tuning. The computation of the cross-correlation matrix in the Barlow module helps a lot in obtaining the latent contrastive loss value. In the test phase, the degraded image is simply passed through the enhancement network to obtain the recovered image. In addition, the two subplots at the bottom right of the figure divide the invariant and redundant components of the matrix. The aim is to maximize the invariant component while minimizing the redundant component to zero.

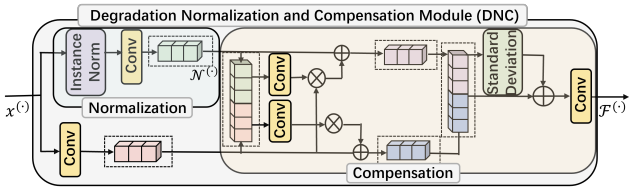


Figure 3: The architecture of DNC (Degradation Normalization and Compensation) module.

channel and spatial dimensions in the projecor section. The reduction is done in a specific order, first the channel dimensions and then the spatial dimensions. This design choice is in line with the UNet's philosophy, as its aim is to minimize information loss at this stage, so as not to interfere with subsequent loss calculations. W_i refers to the parameters of the convolutional layer used to reduce the channel dimension. We repeat this dimension reduction process twice in the reduction module, which can be expressed as:

$$\mathcal{P}_i = pool(relu(W_i * bn(\cdot)), i = 1, 2 \tag{2}$$

We refer to the invariance term of the two latent representations as $I(r^m, r^{m1})$ and the redundancy term as $R(r^m, r^{m1})$. Their visual schematic is shown in the bottom right of Figure 2, and a more detailed computational procedure is described in Algorithm 1. The latent contrast loss is defined as follows:

$$\begin{aligned}\mathcal{L}_{Lcon} &= max\{I(r^m, r^{m1})\} + min\{R(r^m, r^{m1})\} \\ &= \Sigma_i(1 - C_{i,i})^2 + \lambda\Sigma_i\Sigma_{j\neq i}C_{i,j}^2\end{aligned} \tag{3}$$

where λ is a positive constant used to weight the importance of the first and second terms of the loss, and C is the cross-correlation matrix computed between the normalized outputs of the two samples. We keep λ the same as in Barlow Twins.

$$C_{i,j} = \frac{\Sigma_b r_{b,i}^m r_{b,j}^{m1}}{\sqrt{\Sigma_b (r_{b,i}^m)^2}\sqrt{\Sigma_b (r_{b,j}^{m1})^2}} \tag{4}$$

The latent contrast loss consists of two terms, the invariance term and the redundancy reduction term. The invariance term, by enforcing the diagonal elements of the cross-correlation matrix to be 1, makes the embedding robust to distortions. The redundancy reduction term, on the other hand, by enforcing the off-diagonal elements of the cross-correlation matrix to be 0, decorrelates the different dimensions of the embedding. This decorrelation ensures that the sample information conveyed by the output is non-redundant.

Loss Function Formulation

Our model uses a different set of losses at each stage. During pre-training, we use four different losses, $\mathcal{L}_{MSE}$, $\mathcal{L}_{BCE}$, $\mathcal{L}_{Dcon}$ and $\mathcal{L}_{Lcon}$. Two of the loss functions, $\mathcal{L}_{MSE}$ and $\mathcal{L}_{BCE}$ are used for pure image restoration. The former is calculated using mean square error and the latter uses binary cross entropy. The inclusion of loss function $\mathcal{L}_{BCE}$ aims to enhance the clarity of character boundaries. $\mathcal{L}_{Dcon}$ and $\mathcal{L}_{Lcon}$ losses are used to supervise different contrastive learning in direct (pixel level) and latent space, respectively. Thus, the total loss function for the pre-training model is:

$$\mathcal{L}_{total} = \mathcal{L}_{MSE} + \lambda_1\mathcal{L}_{BCE} + \mathcal{L}_{Dcon} + \lambda_2\mathcal{L}_{Lcon} \tag{5}$$

In terms of $\mathcal{L}_{MSE}$, we set its coefficient to 1 since it is similar in magnitude to $\mathcal{L}_{Dcon}$. Regarding λ_1, we set its value to 5 based on hyper-parameter experiments, since our model achieves optimal performance with this settings. $\mathcal{L}_{Lcon}$, on the other hand, is calculated over the whole feature dimension and possesses a significantly larger value compared to other elements of the loss function. Therefore, we set λ_2 to

Algorithm 1: PyTorch-style pseudocode for Contrastive Degradation Representation

```
# net: enhancement net
# proj: project head
# lambda: weight on off-diagonal terms
# N: batch size
# D: dimensionality of the embeddings
# mm: matrix-matrix multiplication
# eye: identity matrix

for x in loader:
   # initialize contrastive loss
   Dcon, Lcon = 0, 0

   for z in degradations:
      # compute embeddings
      e_x, n_x = proj(Norm(x)), net(x)
      e_z, n_z = proj(Norm(z)), net(z)

      # normalize along the batch
      e_x_norm = (e_x - e_x.mean(0)) /
      e_x.std(0)
      e_z_norm = (e_z - e_z.mean(0)) /
      e_z.std(0)

      # cross-correlation matrix
      c = mm(e_x_norm.T, e_z_norm) / N

      # loss
      # direct contrastive loss
      Dcon.append(mse(n_x, n_z))
      # latent contrastive loss
      c_diff = (c - eye(D)).pow(2)
      off_diagnal(c_diff).mul(lambda)
      Lcon.append(c_diff.sum())

   loss = loss_restoration + Dcon + Lcon
   loss.backward()
   optimizer.step()
```

$1/100$ to ensure improved performance and maintain balance between different loss elements.

In the fine-tuning phase, we only use L_{MSE}. Our model has learned good representations of degraded document images, so we can adopt a simpler form of loss function.

Experiments

Datasets

Our experiments are trained and evaluated on a variety of degraded document datasets. These datasets cover a wide range of document types such as handwritten, printed, ancient, and modern documents. The data setup differs in the pre-training and fine-tuning phases. To prevent the model from simply memorizing the data rather than learning from it, we generated a large number of degraded samples for model pre-training using ground truth data from the DIBCO 2008-2019 dataset series. To maintain fairness in the across-dataset test, images from the test set are excluded from the pre-training dataset. Specifically, the training, validation, and test sets consist of 10,937, 100, and 60 pairs of high-resolution complete document images, respectively. During the fine-tuning phase, the model will be fine-tuned using all real data from the DIBCO 2008-2019 dataset series (DIBCO series for short) except the test data, so the amount of training data in each testset is not fixed. Please refer to the supplementary material for more details.

Implementation Details

We train our networks using small image patches randomly sampled from the document images. The base patch size is set to 256×256. So for images with heights less than 256, we resize them to 384 instead. We also use augmentation methods (flipping and rotation) to create more complex training samples. In flipping, we flip patches vertically or horizontally at a 50% ratio. When rotating, we rotate patches by 180 degrees with a 50% ratio.

Our training batch size is set to 16 throughout the training process. In the pre-training phase, the learning rate is initially set to 1e-4 and the number of training iterations is 60000. The system runs on an Ubuntu server platform with two GPUs (NVIDIA GeForce RTX 2080 Ti with 11G memory).

Quantitative Results

For test images, we use four publicly available datasets of independently degraded document images from previous DIBCO competition series. We evaluate not only the generelization ability of the pre-trained models, but also their fine-tuning performance.

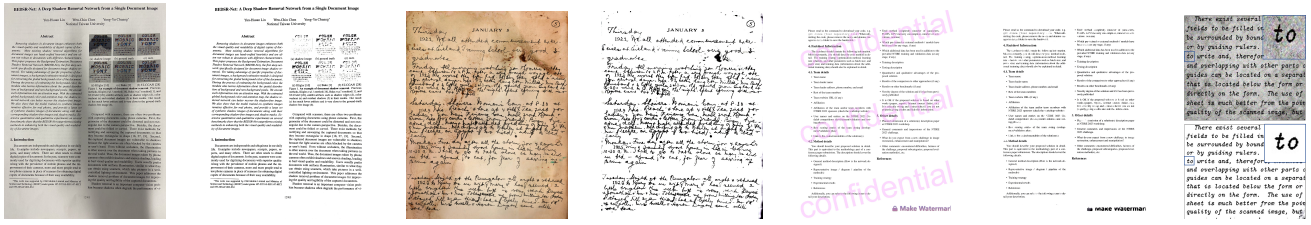
Unified Pre-training Results In the pre-training phase, the baseline method is set to be the official Unet. We adopt the augmented ground truth images as pre-training resources, while the test sets are arbitrary. On the same pre-training and test datasets, we compare the proposed method with three representative heterogeneous state-of-the-art methods, including; a Transformer-based method DocEnTr (Souibgui et al. 2022), a GAN-based method DE-GAN (Souibgui and Kessentini 2020) and a Unet-based method BCDU-Net (Azad et al. 2019), which uses the same backbone enhancement network as ours.

We adopt PSNR (Huynh-Thu and Ghanbari 2008) and SSIM (Wang et al. 2004) as the basic evaluation metrics for image quality assessment to evaluate the image restoration effect: the higher the PSNR and SSIM, the more natural the signal-to-noise ratio and structural similarity, and the better the perception. The PSNR and SSIM results on four datasets are shown in Table 1. In terms of PSNR and SSIM, DocNLC wins across the board, with the best average performance across all four datasets. We attribute the good generalization ability of the network to the robust degradation representations learned by the direct contrastive loss and latent contrastive module. More importantly, DocNLC has the least increase in parameters compared to the baseline model.

Fine-tuning Results Although DocNLC shows good generalization performance for the four testsets in the pre-training phase, previous representative models mostly provide their fine-tuned results. Therefore, we conduct additional experiments to prove that the prior knowledge learned by DocNLC in the pre-training phase can improve its fine-tuned results. For a complete evaluation of the reference

Method	DIBCO11		DIBCO12		DIBCO17		DIBCO18		Average		#Param
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	
U-net (MICCAI 2015)	13.54	0.83	15.40	0.89	12.42	0.82	13.86	0.85	13.81	0.85	29.60
BCDUnet (ICCV 2019)	15.89	0.88	17.06	0.92	15.15	0.86	16.01	0.89	16.03	0.89	80.24
DE-GAN (TPAMI 2020)	13.36	0.81	16.42	0.89	13.94	0.82	12.12	0.79	13.96	0.83	118.39
DocEnTr (ICPR 2022)	15.11	0.85	15.71	0.89	14.72	0.85	15.04	0.86	15.16	0.86	276.4
DocNLC(Ours)	**16.42**	**0.89**	**19.36**	**0.94**	**16.59**	**0.88**	**16.31**	**0.90**	**17.17**	**0.90**	**29.84**

Table 1: Quantitative comparison of the generalization ability of different pre-trained methods in terms of PSNR and SSIM.



(a) Shadow (b) Ours (c) Background (d) Ours (e) Watermark (f) Ours (g) Noise

Figure 4: Visualisation of background, shadow, watermark and noise removal. Document images with shadows, backgrounds and watermarks are shown in (a), (c) and (e). The results after enhancement by pre-trained DocNLC are shown in (b), (d) and (f). In addition, (g) shows the original and the processed image with noise.

image quality, we still use the PSNR and SSIM metrics to quantitatively compare the performance of different methods.

For fair evaluation on the DIBCO series, we mark the entire DIBCO datasets as D_{full} and the dataset for the year we want to test as T, so the training set in this case is $D_{full} - T$. "D" in Table 2 has the same meaning as $D_{full} - T$. If the number of training data sets is less than $D_{full} - T$, we mark this as $D-$. If it is more than that, it is marked as $D+$. As shown in Table 2, DocNLC achieves the SOTA PSNR and SSIM in three out of four test sets and performs well overall. As for the performance on DIBCO2018, we attribute them to the different data distribution compared to other years. In particular, the text colour in DIBCO2018 is more similar to the background colour making the text areas difficult to separate.

Qualitative Results

In order to assess the generalization capabilities of our model over different degraded documents, Fig. 4 illustrates the effect of different degraded document images before and after enhancement. Figure 4 contains four pairs of images, from left to right: an image with a background and its enhancement effect, an image with a shadow and its enhancement effect, an image with a watermark and its enhancement effect, and an image with noise and its enhancement effect. These examples demonstrate that our model can efficiently process a single shading layer, successfully remove the coloured watermarks and recover the foreground characters from noisy backgrounds. The readability of the document image is significantly improved after the enhancement process.

Ablation Study

Loss Function Analysis We perform ablation studies on various loss functions to verify their validity. Note that since both direct and latent contrastive loss are assisted components, we do not report the performance when they are used alone. The image restoration loss $\mathcal{L}_{MSE}$ and $\mathcal{L}_{BCE}$ are the primary loss functions, so we give the results when these two losses are used alone. As shown in Table 3, using the $\mathcal{L}_{BCE}$ loss improves the PSNR and SSIM results compared to the $\mathcal{L}_{MSE}$, which means better image restoration. This may be due to the fact that when using the $\mathcal{L}_{BCE}$, document image enhancement is treated as a two-class segmentation task, which is more difficult and the model is less prone to overfitting. Table 3 shows the PSNR and SSIM results for different losses on four test sets. More visulization results are shown in Figure 5.

The contrastive learning losses $\mathcal{L}_{Dcon}$ and $\mathcal{L}_{Lcon}$ play a significant role in improving the document image quality. The direct contrastive loss $\mathcal{L}_{Dcon}$ mainly focuses on reducing the differences in mean square values between image pixels, while the latent contrastive loss $\mathcal{L}_{Lcon}$ greatly improves the quality of the image as perceived by the human eye. Compared to the results in the last row of Table 3, the absence of any of the contrastive losses leads to a sharp drop in the model's performance on multiple test sets.

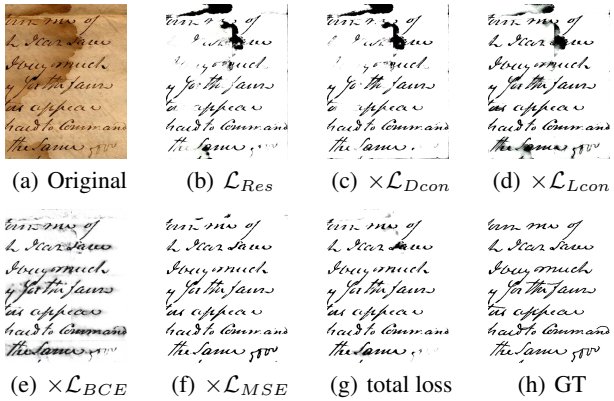
Primary Modality Ablation Results Table 4 shows that training with "background" as the primary modality gives the best results. Intuitively, this is because "background" occurs in the first stage of document degradation, when the mutual information difference between the degraded map and the ground truth is at a minimum with respect to the ground truth itself, and thus the model recovers higher qual-

	Method	Training Datasets		PSNR			
		Pretrain	Finetune	DIBCO11	DIBCO12	DIBCO17	DIBCO18
Rule Based	Otsu (Otsu 1979)	✗	✗	15.7	15.03	14.25	9.74
	Sauvola (Sauvola and Pietikäinen 2000)	✗	✗	15.6	16.71	14.25	13.78
Deep Model	Competition Winner (Pratikakis et al. 2018)	-	-	16.1	21.80	18.28	19.11
	Cascaded cGan (Zhao et al. 2019)	✗	D-	20.30	21.91	17.83	18.37
	DeepOtsu (He and Schomaker 2019)	✗	D+	19.9	-	-	-
	DE-GAN (Souibgui and Kessentini 2020)	✗	D-	-	-	18.74	16.16
	CMU-Net (Kang, Iwana, and Uchida 2021)	COCO-Text	D	19.9	21.37	15.85	19.39
	DocEnTr(Souibgui et al. 2022)	✗	D+	20.81	22.29	19.11	**20.18**
	Text-DIAE (Souibgui et al. 2023)	unknown	unknown	21.29	23.66	19.64	19.95
Ours	DocNLC(Pre-training)	Aug	✗	16.42	19.36	16.59	17.17
	DocNLC(Fine-tuning)	Aug	D	**22.15**	**23.91**	**20.24**	18.23

Table 2: Comparison of PSNR and SSIM performance of different methods on four DIBCO datasets after fine-tuning. D+: DIBCO supersets, D: full set of DIBCO, D-: DIBCO subset

$\mathcal{L}_{MSE}$	$\mathcal{L}_{BCE}$	$\mathcal{L}_{Dcon}$	$\mathcal{L}_{Lcon}$	DIBCO2011		DIBCO2012		DIBCO2017		DIBCO2018		Average	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
✓				13.74	0.80	15.40	0.87	12.78	0.82	14.42	0.85	14.10	0.84
	✓			13.75	0.85	15.69	0.90	13.27	0.85	13.87	0.87	14.15	0.87
	✓	✓	✓	15.17	0.89	17.89	0.93	15.94	**0.89**	16.14	0.90	16.29	0.90
✓		✓	✓	16.07	0.83	18.07	0.91	16.33	0.84	16.54	0.86	16.75	0.86
✓	✓		✓	13.97	0.85	16.49	0.91	13.20	0.84	14.60	0.88	14.57	0.87
✓	✓	✓		15.82	0.86	18.37	0.92	16.41	0.88	**16.64**	0.88	16.81	0.89
✓	✓	✓	✓	**16.42**	**0.89**	**19.36**	**0.94**	**16.59**	0.88	16.31	**0.90**	**17.17**	**0.90**

Table 3: Results of loss function ablation



(a) Original (b) $\mathcal{L}_{Res}$ (c) $\times\mathcal{L}_{Dcon}$ (d) $\times\mathcal{L}_{Lcon}$

(e) $\times\mathcal{L}_{BCE}$ (f) $\times\mathcal{L}_{MSE}$ (g) total loss (h) GT

Figure 5: Ablation visualization of different loss function components. $\mathcal{L}_{Res}$ represents for $\mathcal{L}_{MSE}$ and $\mathcal{L}_{BCE}$. These visualizations align well with the quantitative results presented in Table 3.

Primary Modality	PSNR	SSIM
Blur	13.08	0.82
Noise	16.14	0.82
Watermark	15.80	0.86
Shadow	16.06	0.87
Background	**16.31**	**0.90**

Table 4: Comparison of results for different types of degradation as "primary modality".

ity document images from this type of degradation. A high quality benchmark can only induce other types of degradation to approach it, and a counter-example can be found in the case where "blur" is used as the primary modality. Blur itself occurs at the last stage of image degradation with the least amount of mutual information, so the quality of the recovered images when it is used as the primary modality is not as high as when the other types of degradation are used as the primary modality at the beginning of the training period, and therefore cannot serve as a good basis for comparison. The poor performance of using "blur" as the primary modality coincides with the theory of inequality in data processing. Now, our answer to the question posed at the beginning is that samples of different degradation types and their high-dimensional features can improve the performance of the DIE model, but the primary modality should be judiciously chosen with higher mutual information with the ground truth.

Although both "blur" and "noise" occur in the final stage, "blur" directly destroys the boundaries of the characters. Therefore, taking 'noise' as the main modality gives better performance. The poor performance for blurred document images is one of the limitations of our work. More details

can be found in the supplementary material. All experiments are tested on the DIBCO2018 dataset.

Conclusion and Limitation

In this paper, we present a document image enhancement framework with normalized and contrasted degraded representations (DocNLC). Our framework focuses on establishing consistency between degradation representations by exchanging degradation information both directly and in latent space. Furthermore, we incorporate a fine-tuning strategy to improve the network's performance on specific datasets. Comprehensive experiments on four DIBCO datasets, commonly used to evaluate state-of-the-art models, demonstrate the superiority of our proposed method. However, it is worth noting that our method performs poorly when dealing with heavily blurred document images or those with multi-layered shadows. Under such circumstances, the model results in inconsistent character colours in the corrected images. This aspect presents an interesting avenue for future investigation.

Acknowledgments

This research is supported in part by GD-NSF (No. 2021A1 515011870), National Key Research and Development Program of China (2022YFC3301702, 2022YFC3301703), and NSFC (Grant no. 61936003, 61771199).

References

Anvari, Z.; and Athitsos, V. 2021. A survey on deep learning based document image enhancement. *ArXiv preprint*, abs/2112.02719.

Azad, R.; Asadi-Aghbolaghi, M.; Fathy, M.; and Escalera, S. 2019. Bi-directional ConvLSTM U-Net with densley connected convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 0–0.

Bardes, A.; Ponce, J.; and LeCun, Y. 2022. Variance-invariance-covariance regularization for self-supervised learning. *ICLR, Vicreg*, 1: 2.

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

Cover, T. M. 1999. *Elements of information theory*. John Wiley & Sons.

Das, S.; Ma, K.; Shu, Z.; Samaras, D.; and Shilkrot, R. 2019. DewarpNet: Single-Image Document Unwarping With Stacked 3D and 2D Regression Networks. In *Proceedings of International Conference on Computer Vision*.

Feng, H.; Wang, Y.; Zhou, W.; Deng, J.; and Li, H. 2021. DocTr: Document Image Transformer for Geometric Unwarping and Illumination Correction. In *Proceedings of the 29th ACM International Conference on Multimedia*, 273–281.

Gangeh, M. J.; Plata, M.; Nezhad, H. R. M.; and Duffy, N. P. 2021. End-to-End Unsupervised Document Image Blind Denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 7888–7897.

Grill, J.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P. H.; Buchatskaya, E.; Doersch, C.; Pires, B. Á.; Guo, Z.; Azar, M. G.; Piot, B.; Kavukcuoglu, K.; Munos, R.; and Valko, M. 2020. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. B. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 9726–9735. IEEE.

He, S.; and Schomaker, L. 2019. DeepOtsu: Document enhancement and binarization using iterative deep learning. *Pattern Recognition*, 91: 379–390.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Huang, J.; Liu, Y.; Fu, X.; Zhou, M.; Wang, Y.; Zhao, F.; and Xiong, Z. 2022. Exposure normalization and compensation for multiple-exposure correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6043–6052.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, 1501–1510.

Huynh-Thu, Q.; and Ghanbari, M. 2008. Scope of validity of PSNR in image/video quality assessment. *Electronics letters*, 44(13): 800–801.

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 448–456. pmlr.

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.

Kligler, N.; Katz, S.; and Tal, A. 2018. Document enhancement using visibility detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2374–2382.

Li, X.; Zhang, B.; Liao, J.; and Sander, P. V. 2019. Document rectification and illumination correction using a patch-based CNN. *ACM Transactions on Graphics (TOG)*, 38(6): 1–11.

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.

Lin, Y.-H.; Chen, W.-C.; and Chuang, Y.-Y. 2020. Bedsr-net: A deep shadow removal network from a single document image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12905–12914.

Ljubenović, M.; and Figueiredo, M. A. 2019. Plug-and-play approach to class-adapted blind image deblurring. *International Journal on Document Analysis and Recognition (IJDAR)*, 22(2): 79–97.

Ma, K.; Shu, Z.; Bai, X.; Wang, J.; and Samaras, D. 2018. Docunet: Document image unwarping via a stacked u-net. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4709.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Otsu, N. 1979. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1): 62–66.

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.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

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

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. In *2022 26th International Conference on Pattern Recognition (ICPR)*, 1699–1705. IEEE.

Souibgui, M. A.; Biswas, S.; Mafla, A.; Biten, A. F.; Fornés, A.; Kessentini, Y.; Lladós, J.; Gomez, L.; and Karatzas, D. 2023. Text-DIAE: A Self-Supervised Degradation Invariant Autoencoders for Text Recognition and Document Enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

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*, 44(3): 1180–1191.

Tran, P.; Tran, A.; Phung, Q.; and Hoai, M. 2021. Explore Image Deblurring via Encoded Blur Kernel Space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Xie, G.-W.; Yin, F.; Zhang, X.-Y.; and Liu, C.-L. 2020. Dewarping document image by displacement flow estimation with fully convolutional network. In *Document Analysis Systems: 14th IAPR International Workshop, DAS 2020, Wuhan, China, July 26–29, 2020, Proceedings 14*, 131–144. Springer.

Yifei, Z.; Hao, Z.; Zixing, S.; Koniusz, P.; and King, I. 2023. Spectral Feature Augmentation for Graph Contrastive Learning and Beyond. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

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, H.; Wu, Q.; Yan, J.; Wipf, D.; and Yu, P. S. 2021. From canonical correlation analysis to self-supervised graph neural networks. *Advances in Neural Information Processing Systems*, 34: 76–89.

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.