Abstract

Removing noise from scanned pages is a vital step before their submission to optical character recognition (OCR) system. Most available image denoising methods are supervised where the pairs of noisy/clean pages are required. However, this assumption is rarely met in real settings. Besides, there is no single model that can remove various noise types from documents. Here, we propose a unified end-to-end unsupervised deep learning model, for the first time, that can effectively remove multiple types of noise, including salt & pepper noise, blurred and/or faded text, as well as watermarks from documents at various levels of intensity. We demonstrate that the proposed model significantly improves the quality of scanned images and the OCR of the pages on several test datasets.

1. Introduction

Millions of electronic documents, such as contracts and invoices, are reviewed in the normal course of business in the enterprise. A large percentage of them are scanned documents containing various types of noise, including salt & pepper (S&P) noise, blurred or faded text, watermarks, etc. Noise in documents highly degrades the performance of the optical character recognition (OCR) and their subsequent digitization and analysis. The first step toward automating document analysis is to improve their quality using image processing techniques, such as image denoising and restoration. Most of the literature, places attention on removing noise from pictures [27] (e.g., natural scenes) not text documents. However, these techniques may not be directly applicable due to very different nature of text documents.

In the image restoration problem, a degradation function and noise may both affect the quality of images [9]. Examples are deblurring, defading, and inpainting. If in a special case, there is no degradation function, the problem would be a pure image denoising problem (e.g., S&P noise removal).

The current state-of-the-art (SOTA) solutions for image restoration problem are discriminative models based on convolutional neural networks (CNNs), auto-encoders and their variants such as REDNet (residual encoder decoder network) [21, 8], DnCNN (denoising convolutional neural networks) [32], and RDN (residual dense network) [33, 34]. These solutions can generally be formulated as follows:

$$\arg\min_{\theta} \sum_{j} \sum_{i} L(f(x^j_i; \theta) = \hat{y}^j_i, y^j_i)$$  (1)

where $x^j_i$ refers to a noisy patch extracted from the noisy image $x_i$, $y^j_i$ and $\hat{y}^j_i$ are the target and predicted clean patches, respectively, $f$ and $\theta$ refer to the CNN and its parameters. The main shortcoming of this approach, however, is the requirement for the availability of clean target images/documents, which is hard to address in real-world documents. In the literature, the noisy/clean pairs are usually prepared by adding some synthetic noise to clean patches, respectively. However, the synthetic noise does not completely model noise on real images/documents, and therefore, the performance of the network trained on these synthetic data is sub-optimal and highly degraded on real noisy images/documents [14, 31].

To address the lack of noisy/clean pairs, noise-to-noise (N2N) [17], noise-to-void (N2V) [15], and noise-to-self (N2S) [3, 16] training strategies have been proposed. However, these solutions are based on the assumption that the noise is additive zero-mean, and/or independent between pixels [17, 15, 16]. This only covers a specific kind of denoising problem and, therefore, not directly applicable to general image restoration problems, including defading and deblurring. Furthermore, in N2N approach, at least two noisy instances of the same document page are needed, which are not readily available in real settings.

There are several challenges in the design of an end-to-end solution for document image clean-up: 1) Noisy/clean pairs are not available, and therefore, the standard SOTA solutions based on discriminative models can not be employed. 2) There are various artifacts at different intensity levels (intra-class variation) in documents. 3) We prefer to have a single model based on one architecture, and one training strategy, i.e., a unified solution to address all
noise/degradation problems (blind denoising/restoration) as opposed to individual models trained separately for each noise type. Training multiple models raises the problem of routing a document containing a specific artifact to the right model for image clean-up.

This paper addresses these challenges by introducing an end-to-end unsupervised image blind denoising algorithm that presents a single unified model to remove various noise types, without the requirement of paired noisy/clean pages. The main contributions are listed as follows:

1. We propose a novel unified architecture by integrating deep mixture of experts with a cycle-consistent GAN as the base network. We formulate a novel loss function for the proposed model.
2. To the best of our knowledge, the designed unified model is the first that removes various artifacts, including noise (such as S&P noise), and degradation (e.g., faded and blurred text, or watermark) at various intensity levels (image blind denoising).
3. We trained the model on actual noisy documents (not documents with synthetically added noise) without the requirement of noisy/clean images, evaluated it across several public and in-house document datasets, and demonstrated its excellent performance on real documents containing various artifacts.

2. Related work

2.1. Discriminative methods

Deep learning techniques for image denoising were pioneered by discriminative models and CNN/autoencoder architectures. Dong et al. [6] proposed one of the earliest CNN-based models for image denoising in the application of compression artifact reduction with several stacked convolutional layers. Ever since, various architectures and modifications of CNNs have been proposed to improve image denoising, including skip symmetric connections in a model named residual encoder-decoder (RED-Net) [21, 8], denoising CNN (DnCNN) employing batch normalization and residual learning [32], residual dense networks (RDNs) [33, 34], wavelet CNNs [18], feature attention [1], and dual residual networks (DRNs) [19]. For example, using DRNs [19], the authors redesigned one network for each specific denoising task, including additive Gaussian noise removal, deblurring, dehazing, and raindrop removal. The main issue with discriminative approaches is that they need noisy/clean pairs.

2.2. Generative adversarial methods

Approaches based on generative adversarial networks (GANs) [10, 2] are recently used to alleviate the requirement for noisy/clean pairs in several different ways. E.g., Chen et al. [5] proposed to estimate the noise distribution from smooth parts of noisy patches using a GAN and to generate noise samples. They composed pairs of noisy/clean images by adding the estimated noise samples to clean patches, which are successively used to train a CNN for denoising. They claim that the estimated noise by the GAN is more realistic than the synthetic noise usually added in discriminative approaches. Cha et al. [4] went one step further and estimated noisy patches (instead of noise samples) using a GAN. This alleviated the requirement to have clean images. Having two different noisy instances of the same patch/image, they used N2N training strategy [17] to train a CNN for image denoising. Both previous approaches are merely applicable to situations where the noise is zero-mean, additive, and independent of the clean image, which are strong requirements that make them applicable to only a subset of, but not general image restoration problems.

2.3. Image denoising in documents

Discriminative approaches for document image clean-up include a CNN-based approach for deblurring [23], a U-net [22] based approach replacing the skip connections between the encoder and decoder blocks with alternating convolutional and recurrent layers for efficient feature extraction [20], a two-stage CNN-based approach where the first stage is to classify the type of deblurring and the second stage to remove it [13], and conditional GANs (cGANs) [29, 26], which is a supervised image-to-image translation approach [12]. DE-GAN [26], particularly, is recently proposed based on cGANs with a modified loss function with promising results on binarization, deblurring, and watermark removal. However, all these methods, including DE-GAN need noisy/clean pairs generated by adding the corruption to the clean pages/patches and train one model per artifact type. Sharma et al. [24] proposed document image cleansing based on cycle-consistent GANs [36]. This approach does not require noisy/clean pairs, however, they trained the model using these pairs [24]. Also, they have trained one model for each noise/artifact type, whereas we have trained one single model for all types of artifacts.

3. Methods

3.1. Problem statement

Here, we express the generic problem of document image clean-up in two statements:

**Problem statement 1** - In real settings, there are many clean documents, which are unpaired to available noisy pages. The main question is whether we can take advantage of these clean documents in our solution. We can formulate the problem as follows: having two unpaired sets of documents, one set consisting of noisy documents (X) and the other a collection of clean documents (Y), and knowing that these two sets are unpaired, can we transform one set
to have the style of the other? This problem can be formulated as an unsupervised image-to-image translation. Recently, there have been several proposals for solving this problem and one of the most prominent ones is based on cycle-consistent GANs (or cycle-GANs in short) [36]. In fact, in this solution, what is more relevant to our problem is to transform the style of noisy documents to clean documents such that we remove/restore noise/degradation from these documents while preserving their text contents.

**Problem statement 2** - The noisy documents $X$ may contain several different noise types. Our primary goal is to design a single model that can tackle all these noise types. In our solution, we propose to integrate deep mixture of experts into the cycle-GAN.

### 3.2. Cycle-consistent GANs

Suppose there are two sets of unpaired document images $\{x_i\}_{i=1}^{N_x}$ and $\{y_j\}_{j=1}^{M_y}$ taken from two domains of noisy $\mathcal{X}$ and clean $\mathcal{Y}$ images, respectively. A cycle-GAN [36] consists of two generators: forward ($H$) and backward ($F$) generators, and two adversarial discriminators $D_Y$ and $D_X$. The generators transform the data from one domain to another, i.e., $H: \mathcal{X} \rightarrow \mathcal{Y}$ and $F: \mathcal{Y} \rightarrow \mathcal{X}$. The adversarial discriminators aim to differentiate between the outputs of generators and the real data, i.e., $D_Y$ aims to discriminate between $H(x)$ and $y$, whereas $D_X$ tries to distinguish between $F(y)$ and $x$. The objective function in cycle-GAN is based on two losses: the GAN loss that transforms the image style from one domain to another, and the cycle-consistency loss that preserves the contents of the image.

### 3.3. Deep mixture of experts

In order to address the problem statement 2, i.e., designing a single unified model that removes different types of noise from documents, we propose using a mixture of experts model. A naive approach is to combine individual trained cycle-GANs (for each noise type) with an ensemble learning on top (Figure 1a). This model-level combination of experts results in a very complex model that needs as many cycle-GANs as the number of noise types.

Here, we propose an alternative solution with much less complexity based on deep mixture of experts (deep MoEs) [30]. Deep MoE operates within a single model and treats each channel (of a CNN, for example) as an expert. It extends the standard single layer MoE model to multiple layers within a single CNN. Wang et al. [30] derived the equivalence between gated channels in a convolution layer and the classic mixture of experts. A deep MoE consists of three main components: 1) base convolutional network, 2) shallow embedding network, and 3) multi-headed sparse gating network (refer to Figure 1 in [30] for the architecture of deep MoEs). The optimization function for the deep MoEs is based on minimizing three losses: 1) the loss for the base CNN (e.g., cross-entropy loss), 2) the cross-entropy loss for the shallow embedding network, and 3) the $\ell_1$ loss for the gated networks.

### 3.4. The proposed architecture: integrated cycle-GAN and deep MoE

In order to design a unified model for image clean-up to remove various noise types from documents, we propose to integrate the deep MoE with cycle-GAN as the base model. Figure 1b depicts the sketch of the proposed architecture.

The components of base cycle-GAN in proposed architecture include: two generators, i.e., forward $H$ and backward $F$ generators, as well as two discriminators $D_Y$ and $D_X$. The other components, i.e., an embedder $E$, a classifier $C$, and gating networks $g_H^t = G_H^{1,...,L_H}$ and $g_F^t = G_F^{1,...,L_F}$ ($L_H$ and $L_F$ are the number of convolutional layers in $H$ and $F$ generators, respectively) construct the elements of the deep MoE in the architecture.

### 3.4.1 The pipeline architecture formulation

Our goal is to learn the generator function, which generates the cleaned/restored image $\hat{y} \in \mathcal{Y}$ based on the input noisy image $x \in \mathcal{X}$. Every noisy image is labeled by one type of noise, which is defined by $c_x \in \mathcal{C}$. In our experiments, the types of noise/imperfections are S&P noise, blurred, faded text, or watermarked document. During the pipeline training, we sample a tuple composed of a noisy image and its label $(x, c_x) \in \mathcal{X} \times \mathcal{C}$ and a clean unpaired image $y \in \mathcal{Y}$.

The first step in the pipeline is to acquire the embedding vector $e_x \in \mathbb{R}^d$ by applying the noisy image $x \in \mathcal{X}$ to the embedder, i.e., $e_x = E(x)$. Next, we calculate the gating network outputs $g_H^i = G_H^i(e_x) \in \mathbb{R}^{N_H^i}$ for $i \in \{1,...,L_H\}$ and $g_F^i = G_F^i(e_x) \in \mathbb{R}^{N_F^i}$ for $i \in \{1,...,L_F\}$, where $N_H^i$ refers to the number of output channels in the $i$-th convolutional layer in generator $H$. The embedder network also predicts the label $c_x$ of the input image $x$ based on the embedding vector, i.e., $c_x = C(e_x)$.

The next steps of the pipeline apply the gating networks $g_H^i \text{ and } g_F^i$ to the cycle-GAN generators $H$ and $F$, respectively. This generates the cleaned image $\hat{y} = H(x, g_H^1, ..., g_H^{L_H}) \text{ and the noisy image } \hat{x} = F(\hat{y}, g_F^1, ..., g_F^{L_F})$. In both generators, we replace a standard convolutional layer by an MoE convolutional layer, i.e., we calculate $h_t^{i+1} = \sum_{i=1}^{N_H^i} (g_H^i) \cdot K_t^i \ast h_t^i$, where $h_t^i$ is the output tensor of the $t$-th convolution layer and $h_t^i$ refers to a tensor on its $i$-th channel; $\ast$ is a convolutional operator and $K_t^i$ denotes a kernel for the $t$-th layer and the $i$-th channel.

The discriminators work in the same way as in the original cycle-GAN, i.e., $D_X(x) \in [0, 1]$ and $D_Y(y) \in [0, 1]$.

---

1Our solution does not require target clean images during training. We use some metadata (noise type) to train the embedder.
is formulated based on a combination of losses given in Eq. (3). Next, we formulate a novel loss function in Eq. (2) in order to keep the same function structure as in Fig. 1: (a) A na"ive approach for mixture of experts on cycle-GANs. (b) The architecture of proposed unified model. $h^i$ and $f^i$ refer to output tensors of $i$-th convolutional layers of $H$ and $F$, respectively. Flat rectangles are fully connected layers, while other components, i.e., $D_X$, $D_Y$, $E$, are CNNs.

### 3.4.2 The loss function formulation

First, we define two additional functions to formulate cycle-GAN generators with deep MoE layers:

$$H_{\text{MoE}}(x, x) = H(x, G_H^1(E(x)), \ldots, G_H^L(E(x))), \quad (2)$$

and

$$F_{\text{MoE}}(y, x) = F(y, G_F^1(E(x)), \ldots, G_F^L(E(x))). \quad (3)$$

Note that only the noisy image $x$ is provided to the embedder and gating networks. The same argument appears twice in Eq. (2) in order to keep the same function structure as in Eq. (3). Next, we formulate a novel loss function $L_{\text{GAN}}$ for $H_{\text{MoE}}$, $D_Y$, $X$, $Y$ as follows:

$$L_{\text{GAN}}(H_{\text{MoE}}, D_Y, X, Y) = \mathbb{E}_{y \sim Y}[\log D_Y(y)] + \mathbb{E}_{x \sim X}[1 - \log D_Y(H_{\text{MoE}}(x, x))]. \quad (4)$$

Similarly, we can define the $L_{\text{GAN}}$ for $F_{\text{MoE}}$ as follows:

$$L_{\text{GAN}}(F_{\text{MoE}}, D_X, Y, X) = \mathbb{E}_{x \sim X}[\log D_X(x)] + \mathbb{E}_{y \sim Y}[1 - \log D_X(F_{\text{MoE}}(y, x))]. \quad (5)$$

On the other hand, the cycle-consistency loss $L_{\text{cyc}}$ for $H_{\text{MoE}}$ and $F_{\text{MoE}}$ is defined as provided below:

$$L_{\text{cyc}}(H_{\text{MoE}}, F_{\text{MoE}}, X, Y) = \mathbb{E}_{x \sim X}[^{|| F_{\text{MoE}}(H_{\text{MoE}}(x, x)), x - x ||}_1] + \mathbb{E}_{y \sim Y}[^{|| H_{\text{MoE}}(F_{\text{MoE}}(y, x)), y - y ||}_1]. \quad (6)$$

The final objective function for cycle-GAN network is formulated based on a combination of losses given in Eqs. (4), (5), and (6):

$$L_{\text{cycle-GAN}}(H_{\text{MoE}}, F_{\text{MoE}}, D_X, D_Y, X, Y) = L_{\text{GAN}}(H_{\text{MoE}}, D_Y, X, Y) + L_{\text{GAN}}(F_{\text{MoE}}, D_X, Y, X) + \lambda_{\text{cyc}} L_{\text{cyc}}(H_{\text{MoE}}, F_{\text{MoE}}, X, Y). \quad (7)$$

In addition, we formulate the loss function required for training deep MoE Layers as follows:

$$L_{\text{MoE}}(E, G_H^1, G_F^1, C, X, C) = \mathbb{E}_{x \sim X}[\text{CrossEntropy}(C(E(x, c)))] \quad \quad (8)$$

The total and final objective function for training the integrated cycle-GAN and deep MoE Layers is derived by combining the losses given in Eqs. (7) and (8):

$$L_{\text{cycle-GAN}}(H_{\text{MoE}}, F_{\text{MoE}}, D_X, D_Y, X, Y) + \lambda_{\text{MoE}} L_{\text{MoE}}(E, G_H^1, G_F^1, C, X, C). \quad (9)$$

### 4. Implementation details

#### 4.1. Model architecture

The cycle-GAN network is adopted from the implementation explained in [36]. ResNets of nine blocks are used in the generators. The discriminator networks are $70 \times 70$ Patch-GANs [12], which classify $70 \times 70$ overlapping patches as real or fake. The embedder network is a 7-layer CNN with kernel size of $3 \times 3$, batch normalization and ReLU activation functions. The classifier $C$, which is the last layer of the embedder is a fully connected layer with softmax and cross-entropy loss function for classification of the input patches based on their noise type (4 classes). Finally, the gating networks are fully connected networks with ReLU activation functions, whose inputs are from the penultimate layer of the embedder, i.e., just before the classifier $C$. There are 18 gating networks for the two
generators, 9 for each generator corresponding to the number of blocks per generator. It is worthwhile to highlight here that these networks are essential during training of the model. For the inference, i.e., for document image clean-up, a minimal model is required, including the forward generator $H^{MoE}$, the embedder $E$, and the gating networks $g_H^{t}$ corresponding to the forward generator $H^{MoE}$. This greatly reduces the network complexity during the inference in terms of network latency and memory footprint.

4.2. Hyper-parameters and model training

Our experiments showed that the same as the original cycle-GAN [36], using a least-square loss instead of the negative log likelihood or Wasserstein loss for $L_{GAN}$ results in a more stable training and a better performance in document image clean-up application, and hence we adopted this loss during training the model. Also, same as the original cycle-GAN, as suggested by Shrivastava et al. [25], the discriminators were updated based on a history of 50 previously generated images instead of just the last generated image by the generators. We empirically set $\lambda_{cyc} = 10$, $\lambda_{MoE} = 1$, and $\lambda_{G_{F}} = \lambda_{G_{H}} = 0.1$. Adam optimizer with the batch size of 32 was used with a learning rate of $2 \times 10^{-4}$. We used 4 Nvidia Tesla V100 GPUs for training the model, which took about 1.5 days to complete. The model was implemented using PyTorch framework.

5. Experimental setup and results

5.1. Training dataset

There are three main document types: unstructured (such as lease contracts, and scientific papers), semi structured (e.g., invoices), and structured (like tax forms). In order to prepare a training set, document pages of various types and noise contents were selected from our in-house documents. The most common noise types on lease contracts are S&P noise, blurred, or faded text, whereas tax forms and invoices are mostly in digital format containing watermarks. The number of pages in each category is shown in Table 1 along with other details about the dataset. Overlapping patches of $256 \times 256$ pixels with the stride of 128 pixels were extracted from these pages to train the network.

The set of noisy and clean pages for the lease contracts are completely unpaired. As for the tax forms and invoices, extracting patches of size $256 \times 256$ pixels from the original watermarked pages results in only 10% patches with watermark (since only a small part of the page is watermarked). Therefore, we synthetically added watermarks to the grids of $4 \times 2$ of clean tax forms and invoices with the variations in fonts, text, size, orientation (0, ±45), colors (gray, light gray, red, purple, and blue), and transparency (scale of 0.1 to 0.6), randomly selected from a uniform distribution (a sample page is shown in Figure 4j). Although this approach generates pairs of watermarked/clean pages, we have not used this information in training the proposed model as the patches were randomly selected from the sets of watermarked and clean pages for training the model irrespective of how they were paired.

5.2. Test datasets

In order to assess the performance of trained unified model, three datasets are used.

**Dataset I**: 100 clean and high quality pages from scientific papers. This provides more tightly controlled conditions for quantitative assessment as the OCR on the original high-quality pages can be considered as the ground truth. Watermarks were synthetically added to each page.

**Dataset II**: 100 pages selected from Tobacco800 dataset [35]. These pages were originally noisy and were cleaned using our unified image clean-up approach.

**Dataset III**: A dataset of 300 in-house noisy documents containing various noise types, including S&P noise, blurred/faded text, and watermarks.

5.3. Evaluation metrics

Since the ultimate purpose of image cleansing is improving the performance of the OCR, we used the improvement in OCR as the metric for the quantitative assessment of the model. We used ABBYY FineReader12 as the OCR engine, and OCRed both the original noisy document pages as well as cleaned ones (outputs of the model). Since the ground truth for the characters are not available, we extracted words from characters on each page and used a relative metric: the words found on the cleaned page were considered as the reference and compared with the words on the noisy page. The percentage of mismatches between the two was then computed as a metric to measure the amount of improvement. In case that the original clean pages are available (e.g., in Dataset I), we compared them as a reference with the noisy (watermarked in Dataset I) pages as well as cleaned pages. We then provided the relative metric as a measure of deterioration. Although this is not a perfect metric, it provides a reasonable quantitative assessment in the absence of ground truth for characters. We reported the averaged percentage improvement/deterioration, maximum improvement/deterioration, and the percentage of pages improved/deteriorated more than 5% and 10%.

In addition, since for Dataset I, the original clean and high quality images are available, we have provided the peak signal-to-noise ratio (PSNR) metric as well.

5.4. Ablation study

In order to demonstrate the effectiveness of gating networks, we have provided the visualization of their outputs. In Figure 2a, the Pearson correlation coefficients are calculated between vector outputs of forward gating network...
Table 1: The details of the dataset used for training the proposed model.

| Document Types     | Lease Contracts | Tax Forms | Invoices |
|--------------------|-----------------|-----------|----------|
| Noise Types        | S&P             | Blurred   | Faded    | Watermarked | Watermarked |
| No. of noisy/clean pages | 663/1125       | 1513/1125 | 377/1125 | 5416/5416   | 1339/1339   |
| Page Size / Patch Size | Resized to the Closest Multiple of Patch Size / 256 × 256 |
| Total No. of Noisy Patches | 290,119        | 578,316   | 137,181  | 2,212,729   | 542,121     |
| Total No. of Clean Patches | 440,159        |           |          |             |             |
| Data Augmentation  | Overlapping with the Stride of 128 Pixels |           |          |             |             |

Figure 2: The effectiveness of gating networks on various noise types used in this paper. See text for explanations.

\[ g_{II}^* \] for an input image containing S&P noise (sp1), and another image containing any other noise type (S&P, blurred, faded, or watermarked). The correlations between two samples containing S&P are close to one for all layers, whereas these correlations are much lower between S&P and other noise types. Figure 2b displays 10 consecutive values of a section of gating network \( g_{II}^* \) for the third convolutional layer of forward generator\(^2\). These values are displayed for two samples of every considered noise type. Both these two figures demonstrate that the gating network has similar responses for the same input noise types and different patterns for different noise types. Note that in Figure 2b some values are 0, which is the result of \( \ell_1 \) loss on the gating networks.

### 5.5. Qualitative results

Here, we present sample outputs of the trained model on unseen test patches and pages for various noise and document types\(^3\). Figure 3 depicts samples of cleaned patches (along with the corresponding inputs of noisy patches) for four different artifacts, including S&P noise, blurred, faded text, and watermark. As can be observed from these results, the proposed trained model can efficiently clean patches containing various artifacts at various intensity levels.

Figure 4 displays samples of cleaned pages along with the corresponding inputs of noisy pages. The rightmost images are cleaned pages generated by the proposed approach. The images in the middle are the cleaned pages generated by models trained on individual noise types (one cycle-GAN per noise type), except for the watermarked page where the middle image is cleaned using REDNet [21, 8]. The proposed model is able to remove all noise types without distorting the contents of the page as effectively as individual models solely trained for one noise type. Also, for the watermarked page, the proposed approach was able to remove watermark as good as REDNet, which is a supervised approach and solely trained for watermark removal.

### 5.6. Quantitative results

For dataset I, the OCR on the original high-quality pages are considered as the ground truth, and the OCR on watermarked pages as well as cleaned pages were compared with this ground truth. The results are provided in Table 2. On average, the proposed unified image clean-up improves the performance of ABBYY OCR by approximately a factor of three. The low percentage of relative change in the OCR quality when comparing the original and cleaned pages demonstrates the effectiveness of the proposed algorithm in removing watermarks, and reducing the OCR errors due to watermarks on the page. We compare our proposed approach with REDNet [21, 8] and DE-GAN [26] as two representative discriminative methods. We trained the REDNet and DE-GAN separately using watermarked tax forms as specified in Table 1, and tested them on Dataset I to compare with our results. What we should consider in this comparison is that both REDNet and DE-GAN have solely been trained on watermarked tax forms/clean pairs in a supervised manner, whereas the proposed approach has been trained on all noise types (including S&P, blurred/faded text, and watermark) and all document types in an unsupervised fashion. However, the performance of the proposed method is just slightly inferior to the REDNet and better than DE-GAN. The proposed approach outperforms DE-GAN mainly because DE-GAN, by design, removes color from inputs. Since watermarks on pages can be in color, DE-GAN has more difficulty to remove them. In addition, to demonstrate the effectiveness of the MoE in our proposed approach to train one single model to remove multiple noise types, we compare the proposed approach with a standard cycle-GAN [36, 24] without deep MoE, and also with our

\(^2\)The responses for all layers are provided in the Supplementary.

\(^3\)More results are provided in the Supplementary.
proposed approach without classifier $C$ (both trained the same way as the proposed approach). The results in Table 2 shows that the proposed approach outperforms these two networks as well. This demonstrates that removing the classifier $C$ from the final stage of the embedder makes training the embedder network $E$ less efficient, and deteriorates the overall performance of the model. The classifier $C$ only supports training the embedder network $E$, and this embedding have two important functions: (a) allows the forward generator $H_{MoE}$ to adapt to various noise types (represented by the embedding vectors) and efficiently clean them, (b) informs the backward generator $F_{MoE}$ the type of noise to be added to its input clean image. In the standard cycle-GAN, information about noise is stored in a denoised image. In case of multiple noise types, this confuses the backward generator and the function of embedder is to avoid this confusion.

Neither original clean pages nor the ground truth OCR are available for Datasets II and III. However, the relative OCR performance of ABBYY on the original noisy pages and the corresponding cleaned pages is used to evaluate the
Table 2: The results of quantitative assessment of the proposed model using the *relative metric* and PSNR on Dataset I. The best results, which in this case are lower numbers for OCR-based metrics and higher numbers for PSNR, as *cleaned* documents are compared with the *original clean* ones, are highlighted.

| Measures                      | Original vs Watermarked (%) | Original vs Cleaned (%) | Proposed w/o Classifier | Proposed |
|-------------------------------|-----------------------------|-------------------------|-------------------------|----------|
|                               | RED-Net [8]                 | DE-GAN [26]             | cycle-GAN [24]          |          |
| Averaged Deterioration (%)    | 7.71                        | 3.74                    | 2.54                    | 2.55     |
| Max. Deteriorization (%)      | 53.09                       | 19.70                   | 33.33                   | 14.73    | 9.93    |
| Perc. of Pages +5% Deterioration | 70                          | 14                      | 67                      | 12        | 10      |
| PSNR                          | 35.65                       | 37.43                   | 36.52                   | 37.99     | 38.33   |

Table 3: The results of quantitative assessment of the proposed model using the *relative metric* on Datasets II and III. The best results, which in this case are higher numbers, as *cleaned* documents are compared with *original noisy* ones, are highlighted.

| Measures                      | Dataset II Cleaned vs Original (%) | Dataset III Cleaned vs Original (%) |
|-------------------------------|------------------------------------|-------------------------------------|
|                               | cycle-GAN [24]                     | cycle-GAN (Sequential)              | Proposed | cycle-GAN [24]                     | cycle-GAN (Sequential) | Proposed |
| Averaged Improvement (%)      | 5.94                               | 6.03                                | 7.2       | 6.30                               | 7.10                   | 9.52     |
| Max. Improvement (%)          | 39.02                              | 52.54                               | 63.06     | 48.17                              | 57.98                  | 66.94    |
| Perc. of Pages +5% Improvement | 37                                 | 42                                  | 52        | 52                                 | 58                     | 67       |
| PSNR                          | 22                                 | 24                                  | 28        | 37                                 | 42                     | 49       |

performance of the proposed model. The results are shown in Table 3. Using the proposed approach on average, the OCR is improved by 7.2% and 9.52% on Datasets II and III, respectively. Also, more than 50% of pages on both datasets receive more than 5% improvement in the OCR as a result of cleaning the pages using the proposed unified model. As for comparison with other approaches, there is no *single model* in the literature that can remove all these noise types. Also, discriminative models like REDNet or DE-GAN cannot be used for comparison as they need noisy/clean pairs for training, which are not available here. Nonetheless, We compare the proposed approach with two alternative methods using standard cycle-GAN: 1) same as previous experiment on Dataset I, training one single cycle-GAN (without deep MoE) using a training set consisting of all patches with various noise types. 2) *Standard cycle-GAN with Sequential training*: since it is difficult to train the standard cycle-GAN using all noise types in the training set, we first trained the model on more difficult to learn noise types (e.g., S&P) for several epochs. Then we stopped training and resumed again using patches from noise types that can be removed easier like faded/blurred text. Although this sequential training strategy improved the performance compared to *standard* training, the performance is still far below the performance of individual trained models, as well as the proposed unified model. In addition, *sequential* training is heuristic and needs some trial-and-error in order to find the optimal number of steps, as well as the number of epochs in each step. As can be observed in Table 3, the proposed model significantly outperforms the alternative methods in improving the OCR quality of cleaned pages.

As a final remark, the inference latency of the proposed algorithm was benchmarked on one GPU (Nvidia Tesla V100). On average, it takes 4.49 seconds to clean a page.

6. Conclusion

In this paper, we proposed an end-to-end unsupervised multi-document image blind denoising that presents a unified model to remove various artifacts from all document types, without any need for target paired cleaned pages. We formulated the loss function for the proposed model and demonstrated that it can successfully remove artifacts and improve OCR on various document types. The proposed model is of low complexity and inference latency. In future work, we will replace the embedder and the classifier components by an autoencoder-based approach in order to train the embedding vector in an unsupervised manner.

Disclaimer

The views reflected in this article are the views of the author and do not necessarily reflect the views of the global EY organization or its member firms.
References

[1] Saeed Anwar and Nick Barnes. Real image denoising with feature attention. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3155–3164, 2019. 2

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. 2

[3] Joshua Batson and Loïc Royer. Noise2Self: Blind denoising by self-supervision. In *Proceedings of Machine Learning Research*, pages 2124–2132, 2019. 1

[4] Sungmin Cha, Taeeron Park, Byeongjoon Kim, Jongduk Baek, and Taesup Moon. GAN2GAN: Generative noise learning for blind denoising with single noisy images, 2021. 2

[5] Jingwen Chen, Jiawei Chen, Hongyang Chao, and Ming Yang. Image blind denoising with generative adversarial network based noise modeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3155–3164, 2018. 2

[6] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaou Tang. Compression artifacts reduction by a deep convolutional network. In *IEEE International Conference on Computer Vision (ICCV)*, pages 576–584, 2015. 2

[7] Dheeru Dua and Casey Graff. UCI machine learning repository. Irvine, CA: University of California, School of Information and Computer Science, 2017. 11

[8] Mehrdad J. Gangeh, Sunil R. Tiyyagura, Sridhar V. Pemmaraju, and S. You. editors, *Advances in Neural Information Processing Systems - Volume 27*, pages 6970–6980. Curran Associates, Inc., 2014. MIT Press. 2

[9] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing (3rd Edition).* Prentice-Hall, Inc., USA, 2006. 1

[10] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS’14, pages 2672—2680, Cambridge, MA, USA, 2014. MIT Press. 2

[11] Adam W. Harley, Alex Uikes, and Konstantinos G. Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–995, 2015. 11

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017. 2, 4

[13] Jile Jiao, Jun Sun, and Naqi Satoshi. A convolutional neural network based two-stage document deblurring. In *International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 703–707, 2017. 2

[14] Yoonsik Kim, Jae Woong Soh, Gu Yong Park, and Nam Ik Cho. Transfer learning from synthetic to real-noise denoising with adaptive instance normalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3479–3489, 2020. 1

[15] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void - learning denoising from single noisy images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2124–2132, 2019. 1

[16] Samuli Laine, Tero Karras, Jaakko Lehtinen, and Timo Aila. High-quality self-supervised deep image denoising. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 6970–6980. Curran Associates, Inc., 2019. 1

[17] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2Noise: Learning image restoration without clean data. volume 80 of *Proceedings of Machine Learning Research*, pages 2965–2974, Stockholmsmässan, Stockholm Sweden, 10-15 Jul 2018. PMLR. 1, 2

[18] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 886–895, 2018. 2

[19] Xing Liu, Masanori Suganuma, Zhun Sun, and Takayuki Okatani. Dual residual networks leveraging the potential of paired operations for image restoration. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7000–7009, 2019. 2

[20] Bharat Mamidibathula and Prabir K. Biswas. SVDocNet: Spatially variant U-Net for blind document deblurring. In *Workshop on Document Intelligence at NeurIPS*, 2019. 2

[21] Xiao-Jiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 2802–2810. Curran Associates, Inc., 2016. 2, 6, 11, 17, 18, 19

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, Cham, 2015. Springer International Publishing. 2

[23] Michal Hradiš, Jan Kotera, Pavel Zemčík, and Filip Šroubek. Convolutional neural networks for direct text deblurring. In X. Xie, M. W. Jones, and G. K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 6.1–6.13. BMVA Press, September 2015. 2

[24] Monika Sharma, Abhishek Verma, and Lovekesh Vig. Learning to clean: A GAN perspective. In G. Carneiro and S. You, editors, *Proceedings of Asian Conference on Computer Vision (ACCV) Workshops*, volume 11367 of *Lecture Notes in Computer Science*, pages 174–185, 2018. 2, 6, 8

[25] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2242–2251, 2017. 5
S-1. Architecture of and training the proposed model

Table S1 provides a summary of the networks used in the proposed model (the details are provided in the main paper).

As explained in the main paper, we used in-house documents, including lease contracts, invoices, and tax forms to prepare the training dataset. The most common noise types on lease contracts, which are considered as unstructured documents, are S&P noise, blurred, or faded text, whereas tax forms (structured documents) and invoices (semi-structured documents) mostly contain watermarks.

The set of noisy and clean pages for the lease contracts are completely unpaired. As for the tax forms and invoices, extracting patches of 256 × 256 from the original watermarked pages resulted in only 10% patches with watermark (due to the fact that usually small part of the page is watermarked). Therefore, submitting these patches to the model did not train it adequately for watermark removal. To remedy this problem, we added watermarks similar to what are seen on actual tax forms or invoices, i.e., with the same variations in text, font, size, orientation, transparency, and color to the grids of 4 × 2 of clean tax forms and invoices. This increased the number of watermarked patches by a percentage of more than 60%. A sample page with synthetically added watermark is shown in Figure S8a. The model was trained using the training dataset for 1,700,000 iterations.

S-2. Test sets used for quantitative evaluation

Table S2 provides additional information on the test sets used for the quantitative assessment of the proposed approach as reported in the main paper.

S-3. Additional results

S-3.1. Ablation study

In the main paper, we provided 10 consecutive values of a section of gating network $g_{t+1}$ for the third convolutional layer of forward generator. Here, in Figure S1, we provide these values for eight remaining convolutional layers of forward generator. These values were calculated for two samples of all considered noise types, including S&P noise (blue), faded text (green), blurred text (yellow), and watermarked pages (red). These results are consistent with those displayed for layer three in Figure 2b of the main paper. As can be observed from these plots, there are strong similarities between the values generated by the gating networks for the same noise type, whereas they are different.
Table S1: The architectural details for the proposed model.

| Networks                          | Details on the Components                  | Loss Function                        |
|-----------------------------------|--------------------------------------------|--------------------------------------|
| Generators                        | ResNets (9 blocks)                         | GAN Loss + cycle-consistency loss    |
| Discriminators                    | 70 × 70 Patch-GANs                         |                                      |
| Embedder (MoE)                    | CNN (7 layers, kernel: 3 × 3, batchnorm, ReLU) | Cross-entropy loss                   |
| Classifier (MoE)                  | Fully connected layer with softmax         |                                      |
| Gating Networks (MoE)             | Fully connected (64 × 256), ReLU, 18 of these networks for the two generators) | ℓ₁ loss                               |

Table S2: The details of the test datasets used for quantitative assessment of the proposed model.

| Datasets                  | Dataset I | Dataset II | Dataset III |
|---------------------------|-----------|------------|-------------|
|                           | Scientific Papers | Tobacco800 | Lease Contracts | Tax Forms | Invoices |
| Noise Types               | Watermark | Various    | S&P | Blurred | Faded | Watermark | Watermark |
| No. of Pages              | 100       | 100        | 60  | 100     | 60    | 40        | 40         |

for different noise types. This demonstrates that the gating networks enable the forward generator to process an image in a different way depending on the containing noise type.

To further demonstrate the effectiveness of the gating networks, t-SNE [28] plots for all convolutional layers of forward generator are provided in Figure S2. The plots depict the distributions of gate outputs (256 features) reduced to two main components using t-SNE algorithm. The plots are obtained for 120 document pages containing one of the artifact types, i.e., S&P noise, faded or blurred text, or watermarks (30 pages in each category). The plots for all layers show that the gates outputs are well separated for all four artifact types, which demonstrates the ability of the gating networks to separate the various noise types. From the plots, it can be observed that the least characteristic features are related to blurred pages, as they are sometimes overlapped with the features calculated on faded or watermarked images. On the other hand, it can also be observed that pages containing S&P noise make the most isolated class.

S-3.2. Qualitative results

We have provided more results on a few noisy document pages, including various artifacts, such as S&P noise, faded or blurred text, and watermarks. These results are provided in Figure S3 for a page from Tobacco800 dataset [35], Figures S4 and S5 for two sample pages from CDIP dataset [11], Figure S6 for a few samples from Kaggle dataset [7], and Figures S7, S8, and S9 for an instruction page of a tax form and two pages from a scientific paper with synthetically added watermarks, respectively. In order to demonstrate the effectiveness of the proposed approach in image clean-up, we have compared our approach with the standard cycle-GAN [36] (without integrated deep MoE) trained on multiple noise types, including S&P noise, faded, and blurred pages in Figures S3 and S4. As can be observed from these two figures, the proposed model is much more effective in removing noise without distorting the texts on the pages. Furthermore, in order to demonstrate the improvement in the OCR after removing noise from a page, we have depicted the differences in OCR on part of a page before and after cleansing in Figure S5b. It can be observed that cleansing the page using the proposed approach is quite effective to improve the OCR performance and to generate correct OCR on the cleaned page. It is important to note that the proposed model has not been trained on any samples from Tobacco800, Kaggle, or CDIP datasets. As was explained in Section S-1, the model has only been trained on our in-house documents, including lease contracts, tax forms, and invoices. Nonetheless, it produces excellent noise removal performance across these public datasets as demonstrated in the results depicted on this supplementary note.

For watermark removal problem, we have qualitatively compared the proposed approach with two supervised approaches, i.e., REDNet [21] and DE-GAN [26] in Figures S7, S8, and S9. REDNet and DE-GAN have solely been trained using part of our training dataset containing paired watermarked/clean patches extracted from tax forms, whereas our proposed method has been trained on all noise types (S&P, faded, blurred, and watermarked). As can be observed in Figures S7c, S8c, and S9c, DE-GAN has difficulty to remove watermark from blank parts of the pages and this should be due to the additional loss function the authors have introduced to the model to preserve the text on pages (refer to Eq. (2) and corresponding explanations in [26]). Although our proposed model is unsupervised and has been trained for several noise types, it is as effective as REDNet (a supervised approach solely trained for watermark removal) in removing watermark from pages.
Figure S1: Ten consecutive values of a section of gating network $g_{ij}$ for all convolutional layers of forward generator (except layer three, for which the values are shown in Figure 2b of the main paper).

Figure S2: t-SNE [28] plots for all convolutional layers of forward generator provided for 120 document pages containing S&P noise, faded or blurred text, or watermarks.
Figure S3: Qualitative results on a sample page from Tobacco800 Dataset (a) the whole page, (b) part of the page zoomed in (red box in (a)). The cleaned pages are compared between a standard cycle-GAN [36] and the proposed approach.
Figure S4: Qualitative results on a sample page from CDIP Dataset (a) the whole page, (b) part of the page zoomed in (red box in (a)), (c) part of the page zoomed in (green box in (a)). The cleaned pages are compared between a standard cycle-GAN [36] and the proposed approach.
Figure S5: Qualitative results on a sample page from CDIP Dataset (a) the whole page, (b) part of the page zoomed in (red box in (a)), the words that result in different OCR on the original and cleaned pages are also displayed for comparison.
A new offline handwritten database for
language, which contains full Spanish sen-
tences, has been developed. The Spartacus
database does not contain Spanish sen-
tences, and the Spanish Restricted-domain Task of Cursive
had two main reasons for creating this con-
scant. First, most databases do not contain Spanish sen-
tences, and Spanish is a widespread major language. A
second reason was to create a corpus from seman-
tically rich contexts. These tasks are commonly used in practice
to develop models that go beyond the lexicon
nition process.

As the Cursive database expanded,

A new offline handwritten database for
language, which contains full Spanish sen-
tences, has been developed. The Spartacus data-
base does not contain Spanish sentences, and
the Spanish Restricted-domain Task of Cursive
had two main reasons for creating this con-
scant. First, most databases do not contain Spanish sen-
tences, and Spanish is a widespread major language. A
second reason was to create a corpus from seman-
tically rich contexts. These tasks are commonly used in practice
to develop models that go beyond the lexicon
nition process.

As the Cursive database expanded,

A new offline handwritten database for
language, which contains full Spanish sen-
tences, has been developed. The Spartacus data-
base does not contain Spanish sentences, and
the Spanish Restricted-domain Task of Cursive
had two main reasons for creating this con-
scant. First, most databases do not contain Spanish sen-
tences, and Spanish is a widespread major language. A
second reason was to create a corpus from seman-
tically rich contexts. These tasks are commonly used in practice
to develop models that go beyond the lexicon
nition process.

As the Cursive database expanded,

Figure S6: Qualitative results on a few samples from Kaggle Dataset.
Figure S7: Qualitative results on a sample instruction page of a tax form: (a) watermarked page, and cleaned pages using: (b) RED-Net [21], (c) DE-GAN [26], and (d) proposed approach.
Figure S8: Qualitative results on a scientific paper with synthetically added watermarks: (a) watermarked page, and cleaned pages using: (b) RED-Net [21], (c) DE-GAN [26], and (d) proposed approach.
Figure S9: Qualitative results on a scientific paper with synthetically added watermarks: (a) watermarked page, and cleaned pages using: (b) RED-Net [21], (c) DE-GAN [26], and (d) proposed approach.