

A simultaneous framework for denoising and super resolution for document images

In this chapter, we present the simultaneous framework for super resolution and denoising of document images in reference-based unsupervised setting. This chapter is organized in the following manner: Section 6.1 gives overview and significance of our problem. Section 6.2 briefs about the related work on super resolution performed collectively for document images, followed by unsupervised super resolution techniques. Section 6.3 gives an overview of the original SinGAN model. It then describes our proposed SinGAN-DNSR model and the new loss formulation. Section 6.4 describes the experimental set up and datasets used for evaluating our model. Section 6.5 illustrate the results and their applicability to OCR, and section 6.6 summarizes and concludes this chapter.

6.1 INTRODUCTION

Super-resolution deals with recovering a high resolution (HR) image from a single or a set of low resolution (LR) images. There are various approaches and architectures proposed for Super Resolution (SR), such as CNNs, Laplacian Pyramids and GANs. Supervised training of such architectures requires a large noise-free dataset. Paired LR-HR examples for training are not easily available. In most real world cases, we have access to only the noisy examples. This makes training of supervised models very difficult. For training purposes, it is possible to use synthetic LR images generated using simple and uniform degradation, such as directly down sampling a high-resolution image with bicubic interpolation. Models trained in a supervised way using synthetic LR-HR pairs show limited performance due to their poor generalization ability on real LR images. This is because real LR images differ from synthetic LR images in noise and degradation style.

A number of unsupervised approaches have been developed to overcome the issue of unavailability of paired training data. Unsupervised methods fall short of performance compared to the supervised methods, but still produce reasonable results to make them an effective approach. There are two approaches in the literature which require just the lone input image for training - ZSSR (Shocher et al. [2018]) and SinGAN (Shaham et al. [2019]). These models have been applied to natural images which normally contain abundant information and can lead to effective SR in the unsupervised setting. Natural images contain various textured objects and backgrounds, which have sufficient information to train the different scales of SinGAN for producing the output SR image.

A typical document image depicts less diversity in the foreground content compared to natural images. Direct application of methods which have been successful for natural image super-resolution on a textual image is not effective due to specificities that distinguish textual images from natural ones (Walha et al. [2016]). Text images are composed of several structural primitives like edges, corners, etc., which constitute complex structures such as characters. Text images have regularity in fine patterns which distinguish them from natural images. The difference

of only a few pixels can alter the character completely (e.g. ‘c’ and ‘e’). Poor resolution can compromise the contextual information and affect the readability of the text document. There is a need for a super resolution framework prepared exclusively for textual images. When we super-resolve a noisy text image, it super-resolves the noise also. It is important for the SR module to focus on super-resolving the foreground text while trying to prevent the noise from being super resolved.

In this work we adapt the SinGAN model for our task of simultaneous denoising and super resolution of document images. We call our proposed model the SinGAN-Denoising and Super Resolution model (SinGAN-DNSR). The SinGAN model is used in two steps. In the first step the SinGAN is trained on a clean reference image. The generator learns to generate a clean image and the discriminator learns to distinguish the statistics of the generated (fake) clean image from that of a real clean image. In the second step, the learned discriminator from the first step is used to provide a component of the loss function that plays a crucial role to guide the generator to generate denoised and super-resolved version of the noisy test image. We evaluate our model on public datasets and obtain promising results. We also show that the super resolved and denoised images produced by our model improve the accuracy of downstream OCR of the test image.

Our contributions are two fold:

1. Reference based (one shot) unpaired learning: Our approach uses a single clean reference image to learn the statistics of patches in a clean image and distinguish it from noisy images. We do not require a paired dataset of LR-HR images to train our model. Any clean image can be used as reference. That is, the reference clean image need not be the noisy version of the reference image.
2. We use a novel loss component for training the SinGAN for denoising. The loss component is contributed by a discriminator network trained in adversarial settings in a GAN to generate a clean image. This adversarial training of the discriminator using a reference clean image makes it capable of distinguishing the statistics of a clean image.

6.2 RELATED WORK

In this section we review the past work that has addressed the joint task of denoising (or deblurring) and super resolution using deep networks. These techniques can be broadly classified as supervised or unsupervised. The supervised techniques make use of paired LR and HR images for training. The unsupervised techniques can use training data which may be - (i) unpaired, i.e. the high resolution images do not correspond to the low resolution images (ii) without any target, i.e. only the low resolution images are available for training.

Sharma et al. (Sharma et al. [2017]) proposed a joint denoising and SR framework which is composed of multiple pretrained models that are fine tuned end-to-end. The following 4 models are pretrained - (i) a noisy LR to denoised LR mapping modeled using sparse denoising autoencoder; pretrained with paired examples, (ii) a CNN based autoencoder trained to serve as feature extractor for the denoised LR; pre-trained with paired self-supervision with the target same as the input. (iii) a CNN feature extractor for HR image, modeled similar to (ii). Lastly (iv), a CNN model to map LR features to HR features; pre-trained using paired examples. Models (ii), (iv), (iii) are cascaded and fine tuned end-to-end using LR, HR paired examples. Finally models (i)–(ii,iv,iii) are cascaded and fine tuned end-to-end on noisy LR-HR paired examples in a supervised way. The performance of the system is assessed by examining the improvement in the OCR performance. The two pre-trained stages for denoising and super resolution are cascaded and fine-tuned end to end.

Xu et al. (Xu et al. [2017]) also used GANs for super resolving blurry face and text images. Separate discriminators trained for face and text images contribute to class-specific adversarial loss. The generator uses a deep CNN for generating the deblurred output. The discriminator uses convolution layers followed by a fully connected decision layer. The losses used are adversarial loss, pixel wise loss and feature matching loss. The feature matching loss (Salimans et al. [2016]) makes use of convolutional features extracted from the discriminator. Authors also use a triplet loss in the standard GAN formulation to ensure that the real samples are closer to one another than generated ones.

Ray et al. (Ray et al. [2019]) trained separate modules for document enhancement and recognition. The combined network is jointly optimized. Unlike their previous work which used convolutional autoencoder for denoising, they used a GAN trained with adversarial loss and content loss. The generator of the GAN used for denoising is a REDnet (Mao et al. [2016]) convolutional autoencoder with skip connections. The super resolution module is another GAN trained with adversarial loss, content loss (mean square error) and cycle consistency loss. The generator of this GAN is a deep back projection network (DBPN) (Haris et al. [2018]). The text recognition module is a bi-directional LSTM which takes as input the features from the last up-projection layer of the DBPN.

Peng and Wang (Peng and Wang [2020]) applied a GAN for super resolution of text images. The generator is a convolutional encoder-decoder with skip connections and a spatial attention layer. Pre-training of the generator is done using paired training examples to compute a multiscale structural similarity loss (Wang et al. [2003]), which encourages the GAN to preserve the structural properties of text. The discriminator uses ResNet blocks followed by 3 fully connected layers. In the unpaired training phase, the loss function used for the GAN is a combination of Least Squares-GAN (Mao et al. [2017]) and Relativistic-GAN (Jolicœur-Martineau [2018]) to avoid saturation and to have stability in learning, respectively.

Yamanaka et al. (Yamanaka et al. [2017]) proposed a CNN-based model trained with paired supervision. The network was designed to be computationally frugal with fewer layers. The feature extractor comprised 7 convolutional blocks in cascade, each layer feeding to a reconstruction network. Instead of using an upsampled LR image, the feature extraction network used the given LR image as is for input. The reconstruction module achieved computational efficiency by using network-in-network structure. The reconstruction module produces 4 feature maps; each map represents a corner-pixel of the up-sampled spatial region of the input pixel. Thus, these feature maps produce a 4x resolution output, which is considered as the residue to be added to the bi-cubic upsampled original input image.

Quan et al. (Quan et al. [2020]) proposed an architecture for progressive upsampling and deblurring of document images by cascading a number of modules. Each module extracts feature maps and uses them to predict the improved (deblurred) image. The input low resolution image is used as a reference input to extract the feature maps from the deblurred image from the previous module. The feature maps carry the high frequency details required for deblurring. The input low resolution image is also used as a reference input to predict the improved (deblurred) image using the feature maps.

There are two unsupervised models for super resolution - ZSSR and SinGAN, both developed for natural images. In (Shocher et al. [2018]) proposed the Zero-shot Super-Resolution (ZSSR) model which uses only the given image for learning the mapping from LR to HR. The network is trained to learn the internal image statistics of the given image to generate the SR image. Several versions of the test image are created using data augmentation and they are downsampled to obtain several versions of LR images. The pairs of LR-HR images are used to train the network

for learning the mapping from LR to HR. Shaham et al. (Shaham et al. [2019]) proposed the SinGAN model which trains a pyramid of GANs to capture the global and local information from images. The GANs operating on multiscaled inputs learn the patch characteristics of the input image at different scales. The GANs at the coarser levels learn the global structure and the those at the finer levels learn the local structures with texture information. After training, the SinGAN model can generate diverse, high quality samples carrying the same visual content as the input image. In this model we adapt the SinGAN architecture for document images. With our proposed network induced loss for penalizing noise, we are able to apply the model for simultaneous noise removal and resolution enhancement.

To boost the OCR accuracy, we propose a generative adversarial network (GAN) based framework, where a SR image generator and a document image quality discriminator are constructed. To obtain high quality SR document image, multiple losses are designed to encourage the generator to learn the structural properties of texts. Meanwhile, the quality discriminator is trained based on a relativistic loss function.

6.3 METHODOLOGY

We first describe SinGAN which is the backbone of our model and then explain our proposed DNSR model for document images.

6.3.1 SinGAN

The SinGAN (Shaham et al. [2019]) model is comprised of a pyramid of GAN networks. Each scale n has a patch generator G_n and a patch discriminator D_n . The scale n varies from 0 (finest scale) to N (coarsest scale). Progressive reconstruction begins from the coarsest scale N . The output of generator G_{n+1} is the reconstruction at scale $n+1$, denoted as $(\tilde{x}_{n+1})$. This output is upsampled by a pyramid scale factor r to give $(\tilde{x}_{n+1}) \uparrow^r$. The generator G_n at the next finer scale $n < N$ receives an input g_n comprising $(\tilde{x}_{n+1}) \uparrow^r$ and a noise component z_n , and generates the reconstruction $\tilde{x}_n = G_n(g_n)$ where $g_n = (z_n, (\tilde{x}_{n+1}) \uparrow^r)$. At the coarsest scale N the reconstruction is done using only random noise z_N , to give $\tilde{x}_N = G_N(z_N)$. That is, at scale N , we use $g_N = z_N$, where z_N is the white Gaussian noise at the coarsest scale.

To use the trained SinGAN network for inference we require only the finest scale generator G_0 . The given LR test image is upsampled by scale factor r , and along with noise z_0 , is given as input to the finest scale generator G_0 . The generated output is again upsampled by scale r and fed back as input for the next iteration of reconstruction. The process is repeated k times where k is the number of pyramid stages (scales) in SinGAN. We have $r = \sqrt[k]{s}$, where s is the factor by which the resolution needs to be increased.

Training samples are patches of a single image which is the input to be super-resolved. To be able to generate natural images, the system should be trained to capture the statistics of complex image structures at many different scales. The generative structure consists of a hierarchy of Patch-GANs. Each Patch-GAN captures the patch distribution at a different scale. The GANs have a limited capacity with 5 conv blocks of the form Conv(3x3), each block performing convolution, batch normalization and leaky relu. These blocks learn the residual which is added to the input to generate the output. The coarsest scale has 32 kernels per block. The count is increased by a factor of 2 every 4 scales. Each of the discriminators also have similar 5 conv blocks with 3x3 receptive fields, batch norm and leaky ReLU. The final discrimination score is the average over the patches. The adversarial component of the loss function is the WGAN loss.

To ensure good quality, SinGAN uses a second component of the loss function - the

reconstruction loss. It computes the L2 norm between the reconstructed image at scale n and the corresponding downsampled version x_n of the input.

6.3.2 SinGAN-DNSR

We adapt the SinGAN for simultaneous denoising and super resolution of text images. Our proposed method does not require any supervision in terms of paired examples, but makes use of a reference clean image. The reference clean image is not the cleaned version of the input image. It can be any clean image of any other text in the same language.

Our methodology is divided into two steps: (i) learning the patch statistics of the clean (reference) image and (ii) generating a super resolved and cleaned version of the test image. An overview of our methodology is shown in figure 6.1.

1. Step 1: Learning the patch statistics of the clean (reference) image. In this step we train the SinGAN model for a reference clean image. This allows the discriminators at the different scales of the SinGAN to learn to distinguish real and fake (generated) patches of a clean image at those scales. During training, these discriminators acquire the competence to distinguish the texture pattern in a generated clean patch and the ground truth clean patch at multiple scales. We denote the trained discriminator at scale n as D_n^{ref} . We utilize the trained discriminators at D_n^{ref} for $n = 1toN$ in the following step for denoising and super-resolving a given test image.
2. Step 2: Denoising and super-resolution of a test image. Aligned with our objective of zero-shot denoising and super resolution, we deploy another SinGAN, which we refer to as SinGAN-DNSR, to train on the noisy test image and learn to generate its denoised and super-resolved version. To assist the training of this SinGAN, we use the discriminators D_n^{ref} which were trained in the previous step. The discriminators D_n^{ref} provide a noise penalty that is high if the patches generated by SinGAN are noisy and is low if the generated patches show statistics resembling that of the clean reference image.

SinGAN for processing the input test images trains a set of patch generators G_n and patch discriminators D_n at multiple scales $n = 1toN$. The training phase is guided by multiple loss components - WGAN adversarial loss, style transfer loss, VGG perceptual loss, and a noise penalty.

During training, the Generator network maximizes the objective $D_n^{SR}(G(z_n, (\tilde{x}_{n+1}) \uparrow^r)) + D_n^{clean}(G(z_n, (\tilde{x}_{n+1}) \uparrow^r))$ and weight updates are applied only to the D_n^{SR} components, as we have frozen the weights of D_n^{clean} . The discriminator D_n maximizes the objective $D_n^{SR}(x_n) + D_n^{SR}(G(z_n, (\tilde{x}_{n+1}) \uparrow^r))$.

6.3.3 Loss formulation

6.3.3.1 Feature matching loss:

Inspired from (Xu et al. [2017]), we use feature matching loss to ensure both sharpness and fidelity in the results. We use the VGG perceptual loss formulation. It employs the pre-trained VGG19 model to extract features of the noisy test image and the reference clean image.

The feature matching loss is formulated as the squared L2 norm of the difference between the VGG19 features extracted for the ground truth high resolution image I^{HR} and the super resolved image $G(I^{LR})$ generated by the SinGAN:

$$L_{FM} = ||\phi(I^{HR}) - \phi(G(I^{LR}))||_2^2 \quad (6.1)$$

Here, L_{FM} denotes feature matching loss, ϕ signifies features obtained from VGG19 pretrained

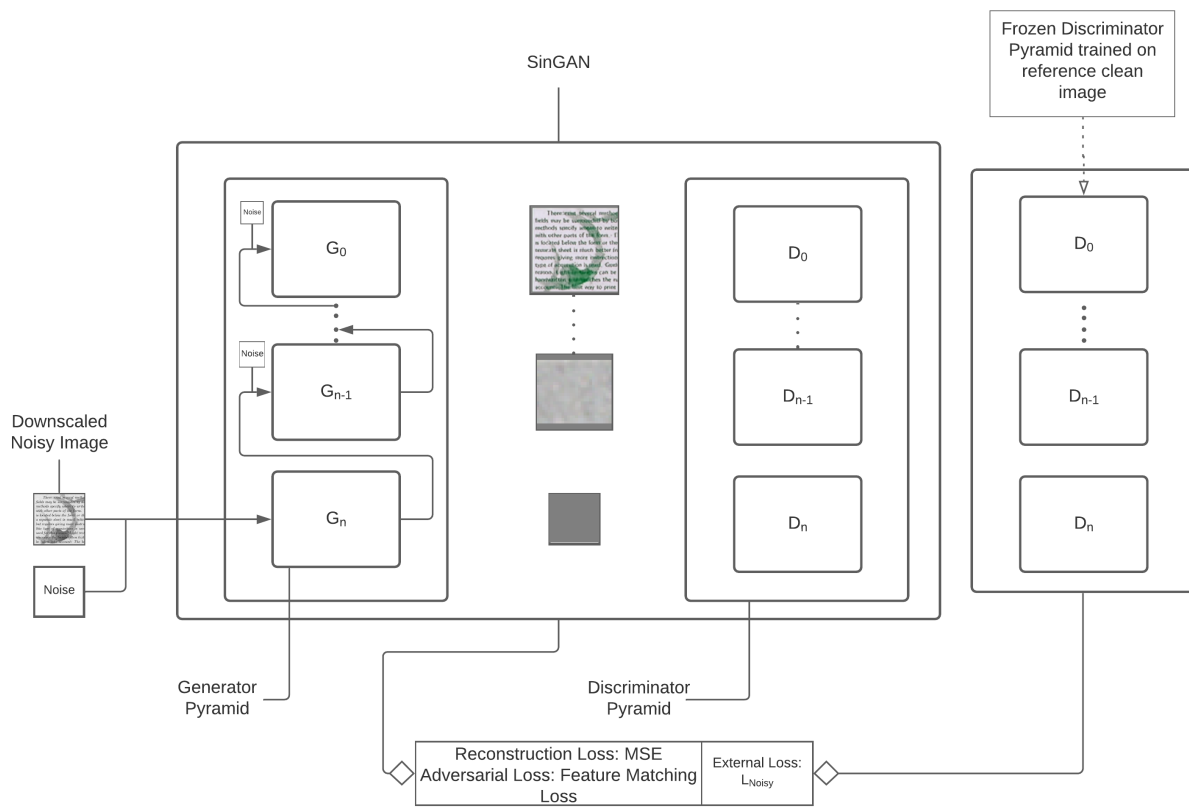


Figure 6.1 : Overview of our methodology

model, and G denotes the generator which super resolves the low resolution image L^{LR} .

6.3.3.2 Style transfer loss:

Inspired from (Johnson et al. [2016]), we have incorporated the style transfer loss in our model. Style transfer works by separating the style and content of the image. Our aim is to copy the style of the desired clean reference image while retaining the content of our noisy test image. Document images may have light or white background with dark or black coloured text as foreground. This style is transferred from the reference clean image to the test image (a text image with noise) while preserving its content. To compute the style transfer loss, we input the test (noisy) image to the VGG19 model and compute the activations of the j^{th} layer of the network. We denote these activations as $\phi_j(I_{noisy})$. Likewise we compute the activations $\phi_j(I_{ref})$ when the clean reference image is given as input to the same model. Assuming that the shape of the feature (activation) map is $C_j \times H_j \times W_j$ we can interpret the feature map as C_j dimensional feature vector defined at each position of the rectangular grid with height H_j and width W_j . For an input x we compute the Gram matrix $G_j^\phi(x)$ defined as:

$$G_j^\phi(x)_{c,c'} = \frac{1}{C_j H_j W_j \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} \phi_j(x)_{h,w,c} \phi_j(x)_{h,w,c'}} \quad (6.2)$$

where h, w are the indices for height, width and c, c' are the indices for the channels. The Gram matrix is a covariance matrix of size $C_j \times C_j$. An entry l, m of the Gram matrix gives the uncentered covariance of the l^{th} and m^{th} components of the feature vector. The summation in Eq 6.2 is over all the grid locations, treating features at each grid location as an independent observation. Highly correlated features tend to co-occur. Hence, the gram matrix captures the style signature of the input.

The style reconstruction loss is computed as:

$$\|G_j^\phi(I_{ref}) - G_j^\phi(I_{noisy})\|_F^2 \quad (6.3)$$

where $\|\cdot\|_F^2$ denotes the squared Frobenius norm and the function $G_j^\phi(\cdot)$ is defined in Eq 6.2.

6.3.3.3 Adversarial loss:

The traditional GAN loss is defined by formulating a minimax optimization for the generator and the discriminator as follows.

$\max_D \min_G \log D(x_n) + \log(1 - D(G(g_n)))$. However, this objective function corresponds to the minimization of the Jensen-Shannon divergence between the distribution of real data $\mathbb{P}_R$ (high resolution images x_n) and the distribution of generated samples $\mathbb{P}_G$ (super resolved images $G(g_n)$). However, this divergence suffers from the problem of mode dropping and instability of gradients.

In this work we use the formulation of WGAN (Arjovsky et al. [2017]; Adler and Lunz [2018]) which is based on minimizing the Wasserstein distance between $\mathbb{P}_R$ and $\mathbb{P}_G$. It is the minimum cost of transporting the probability mass of distribution $\mathbb{P}_R$ from its support to match the probability mass of the distribution $\mathbb{P}_G$ on $\mathbb{P}_G$'s support. The Wasserstein metric incorporates the notion of distance between the elements in the support of $\mathbb{P}_R$ and the ones in the support of $\mathbb{P}_G$. As the support of $\mathbb{P}_R$ and $\mathbb{P}_G$ comes closer, the distance between the manifolds $\mathbb{P}_R$ and $\mathbb{P}_G$ approaches 0.

To be specific we denote the discriminator at scale n with parameters w as $D_{n,w}$ and the generator at scale n with parameters θ as $G_{n,\theta}$. As given in (Arjovsky et al. [2017]) the discriminator

$D_{n,w}$ computes the Wassertein distance by performing the following maximization with respect to its parameters w

$$\min_{w \in W} \mathbb{E}_{z_n \sim p(z_n)} [D_{n,w}(G_{n,\theta}(g_n))] - \mathbb{E}_{x \sim \mathbb{P}_r} [D_{n,w}(x_n)] \quad (6.4)$$

Here the noise z_n is used in the generator input g_n as explained in Section 6.1. Likewise, the adversarial training of the generator updates its parameter θ by minimizing

$$\min_{\theta} -\mathbb{E}_{z_n \sim \tilde{p}(z_n)} [D_{n,w}(G_{n,\theta}(g_n))] \quad (6.5)$$

The parameters w are constrained to lie on compact space, say a fixed box $W = [-0.001, 0.001]^l$ in an l dimensional parameter space. The generator tries to fool the discriminator to achieve high value of the output when fed with fake/generated patches. The discriminator tries to maximize its output when fed with real patches. For stability of training we also add a gradient penalty to the WGAN loss in Equation 6.4 by adding

$$\lambda \mathbb{E}_{\hat{x} \sim \mathbb{P}_{\hat{x}}} [(\|\Delta_{\hat{x}} D_{n,w}(\hat{x})\|_2 - 1)^2] \quad (6.6)$$

The terms used in Equation 6.4 and 6.5 constitute the adversarial loss L_{advers} .

6.3.3.4 Penalty term on noisy patches:

In order to achieve simultaneous denoising and super resolution, we need to impose a penalty when a generated patch contains noise. At scale n we use the discriminator network D_n^{ref} that was trained on the reference image to compute the penalty value for the generated patch. The discriminator D_n^{ref} can distinguish between a generated noisy patch and a clean patch at scale n . It produces an output close to 1 if the generated patch $G_{\theta}(g_n)$ has similar statistics as that of the reference clean patch. We want the loss to be low when the generated patch has less noise. We formulate the loss L_{noise} as follows:

$$L_{noise} = 1 - D_n^{ref}(G_{n,\theta}(g_n)) \quad (6.7)$$

which penalises the noise in the input image. Note that the discriminator network D_n^{clean} does not update its weight when training the SinGAN-DNSR to process the test image. The trained network D_n^{clean} is used to compute the loss using the network function with the generated patch at scale n provided as its input.

6.3.3.5 Reconstruction loss:

The reconstruction loss for scale n is formulated as

$$L_{reconst} = \|G_n(g_n) - x_n\|^2. \quad (6.8)$$

This loss gives the squared error between the generated patch at scale n and the ground truth patch at scale n . This loss ensures that the generated image has the same text content as the noisy image.

6.3.3.6 Total loss:

The total loss to be used for training the SinGAN on the clean reference image is computed as:

$$L_{adver} + \gamma(L_{percep} + L_{styletrans}) + \alpha * L_{reconst}. \quad (6.9)$$

Parameters	Kaggle Dirty Document	DIBCO 2019
Total images in Dataset	144 pairs	20 pairs
Total patches of size 100100	2393	633

Table 6.1 : Statistics of Kaggle Dirty document and DIBCO 2019 dataset

The total loss to be used for training the SinGAN on the noisy test image is computed as:

$$L_{adver} + \gamma(L_{percep} + L_{styletrans}) + \alpha * L_{reconst} + \beta * L_{noisy} \quad (6.10)$$

The component loss terms in the total loss have been defined in equations 6.1, 6.3, 6.4, 6.5, 6.6, and 6.7. We have used weighing parameters γ , α and β , whose optimal values were found to be 1.5, 20 and 0.7 respectively, using cross validation.

6.4 EXPERIMENT

6.4.1 Implementation details

Our SinGAN-DNSR model is adapted from SinGAN with few modifications in parameters to obtain best results for document images. We train the model for 1000 iterations in Step 1 using a clean-LR image. However, in Step 2 while performing super-resolution it is trained for only 100 iterations since we got good performance with fewer iterations. Kernel size for the CNN layers in the Generator is set to 2 in our setting which enables us to capture the local fine details of image. Other parameters and model configuration are the same as those in the original SinGAN model. We use Structural Similarity Index (SSIM) and Peak Signal to Noise Ratio (PSNR) to evaluate the performance of our model. We have also evaluated our model for OCR by using the Pytesseract software. Here, we use confidence value of recognized words as the metric to assess the impact of denoising on text recognition.

6.4.2 Dataset

We have used the Kaggle dirty document dataset, and the DIBCO dataset (Hradiš et al. [2015]) to train/evaluate our model. The Kaggle dirty document dataset (Frank [2010]) is a part of UCI Machine Learning Repository (Dua and Graff [2017]), converted for Kaggle competitions. It comprises of 144 noisy-clean pairs. We downsample the noisy images to get the low resolution (LR) samples. The typical noise in this data resembles coffee stains, faded sun spots, dog-eared pages, and lot of wrinkles which are commonly present when printed documents are stored offline for long time.

The DIBCO-2019 dataset (Hradiš et al. [2015]) was originally proposed for the Document Image Binarization Competition in conjunction with the International Conference on Document Analysis and Recognition (ICDAR) in 2009. This dataset is comprised of machine-printed and handwritten document images. Recently, this dataset is also used for document noise removal task in (Do et al. [2019]; Habibunnisha et al. [2019]; Dumpala et al. [2019]; Souibgui and Kessentini [2020]). The noise in this dataset is background texture noise and ink bleed. Table 6.1 shows the statistics of the Kaggle Dirty Document and DIBCO 2019 datasets.

Model name	Kaggle Dirty Document	DIBCO
ZSSR	19.43/0.88	19.87/0.91
SinGAN	20.13/0.93	21.43/0.94

Table 6.2 : Comparison of results (PSNR and SSIM) for ZSSR and SinGAN for clean images of Kaggle dirty document and DIBCO dataset

6.4.3 Ablation Study

There are two existing state-of-the-art techniques for zero-shot super resolution, however they were proposed for natural images. So, in order to check their applicability on text images, we analyze the performance of these models for super-resolving clean LR images with text content. It was observed that the SinGAN approach performs better in comparison to the ZSSR approach. Table 6.2 shows the performance of ZSSR and SinGAN; for clean-LR Kaggle dirty document and DIBCO dataset images. Table 6.3 provides the qualitative results for ZSSR and SinGAN on clean-LR Kaggle images with their PSNR/SSIM values.

Next, we conducted experiments to analyse the effect of the kernel size and feature loss function on our model. The original architecture uses a kernel size of 3×3 for 5 layers in series. This resulted in overall receptive field of 11×11 . According to our experiments, a receptive field with 11×11 size diluted the effects of minute structures found in shapes of letters in the document. To capture properties of these minute structures we decreased the individual kernel size to 2×2 and hence the effective receptive field reduces to 6×6 . This resulted in preservation of much finer local details. The original model uses mean square error loss along with the adversarial loss. We add the new feature matching loss and style transfer loss along with MSE and adversarial loss. There was a slight improvement in the results after incorporating feature matching and style transfer loss in the loss function. Table 6.4 give results for PSNR and SSIM values for different settings of the model, i.e., the SinGAN-DNSR model with kernel size 3×3 , SinGAN-DNSR model without using feature matching loss, final SinGAN-DNSR model with new loss and kernel size 2×2 applied on Kaggle dirty document dataset and DIBCO image dataset. Table 6.5 provides qualitative results corresponding to table 6.4 for the Kaggle dirty document dataset while table 6.6 shows qualitative result for the DIBCO dataset corresponding to table 6.4.

We have also experimented with different combinations of loss components. We designate Loss 1 as the adversarial loss and reconstruction loss, This formulation was present in the original SinGAN model. We designate Loss 2 as comprising the adversarial, reconstruction as well as the perceptual loss. We designate Loss 3 as the addition of style transfer loss to Loss 2. The effects of these loss function is evaluated in this model. Table 6.7 gives quantitative results for the Kaggle Dirty Document and the DIBCO 2019 datasets for different loss settings. Using perceptual and style transfer loss improves the results in comparison to the original SinGAN loss formulation. Tables 6.8 and 6.9 give qualitative results for various loss function combinations for thr Kaggle Dirty Document dataset and the DIBCO dataset, respectively.

6.4.3.1 Effect of reference clean image:

We experimented with two different reference images to understand its effect on the super resolution task. The first image belongs to a printed Bangla document image and the second image belongs to a natural image. These two images are shown in table 6.10. Table 6.11 shows the generated SR images obtained after using the two reference images. Since the two reference images

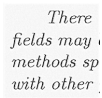
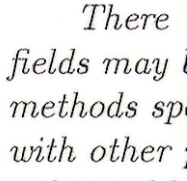
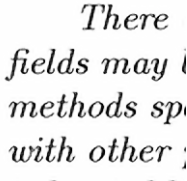
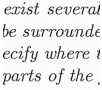
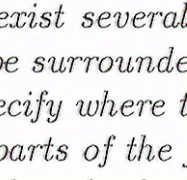
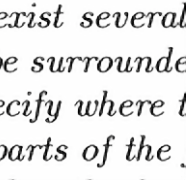
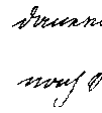
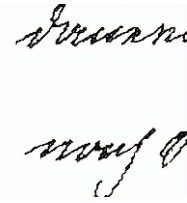
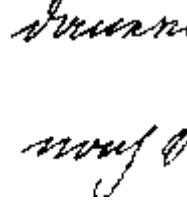
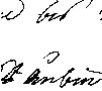
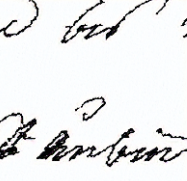
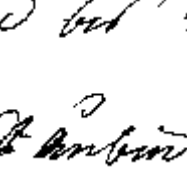
Input	ZSSR	SinGAN
 Kaggle dataset	 19.67/0.86	 20.04/0.89
 Kaggle dataset	 18.97/0.89	 20.01/0.87
 DIBCO dataset	 20.14/0.91	 21.22/0.93
 DIBCO dataset	 20.04/0.89	 20.67/0.91

Table 6.3 : Qualitative results for ZSSR and SinGAN SR for clean images from the Kaggle Dirty Document and DIBCO datasets with their PSNR/SSIM values

Model Name	Kaggle Dirty Document Dataset	DIBCO Image Dataset
	PSNR/SSIM	PSNR/SSIM
SinGAN-DNSR with kernel size 3×3	11.72/0.48	9.67/0.32
SinGAN-DNSR without feature matching loss	12.18/0.53	12.38/0.52
Final SinGAN-DNSR	14.64/0.67	13.14/0.59

Table 6.4 : Ablation experiment on Kaggle dirty document dataset and DIBCO image dataset: PSNR and SSIM values for different settings of DNSR

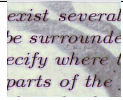
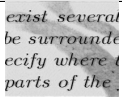
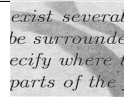
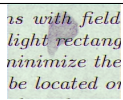
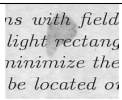
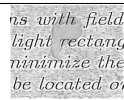
Input	Ground Truth	SinGAN-DNSR1	SinGAN-DNSR2	SinGAN-DNSR3
	<i>exist several be surrounded ecify where t parts of the</i>	 11.53/0.52	 12.52/0.58	 12.99/0.67
	<i>ns with field light rectang minimize the be located on</i>	 11.52/0.46	 13.82/0.52	 15.70/0.71

Table 6.5 : Qualitative results for ablation experiment on different settings of SinGAN-DNSR for Kaggle Dirty document dataset along with their PSNR and SSIM values. Here, SinGAN-DNSR1 corresponds to SinGAN-DNSR with kernel size 3×3 , SinGAN-DNSR2 corresponds to SinGAN-DNSR without feature matching loss function and SinGAN-DNSR3 corresponds to SinGAN-DNSR with feature matching loss function

Input	Ground Truth	SinGAN-DNSR1	SinGAN-DNSR2	SinGAN-DNSR3
		 8.97/0.56	 12.19/0.63	 14.36/0.72
		 9.9/0.68	 8.9/0.60	 15.28/0.8

Table 6.6 : Qualitative results for ablation experiment on different settings of SinGAN-DNSR for DIBCO document dataset along with their PSNR and SSIM values. Here, SinGAN-DNSR1 corresponds to SinGAN-DNSR with kernel size 3×3 , SinGAN-DNSR2 corresponds to SinGAN-DNSR without feature matching loss function and SinGAN-DNSR3 corresponds to SinGAN-DNSR with feature matching loss function

Model Name	Kaggle Dirty Document Dataset	DIBCO Image Dataset
	PSNR/SSIM	PSNR/SSIM
SinGAN-DNSR with original loss	12.18/0.53	9.69/0.65
SinGAN-DNSR with perceptual loss	13.926/0.73	10.35/0.71
SinGAN-DNSR perceptual and style transfer loss	14.33/0.75	11.61/0.77

Table 6.7 : Ablation experiment on Kaggle dirty document dataset and DIBCO image dataset: PSNR and SSIM values for different settings of loss functions

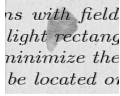
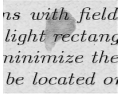
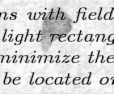
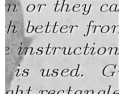
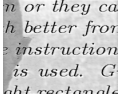
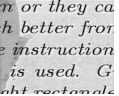
Input	Ground Truth	SinGAN-DNSR Loss1	SinGAN-DNSR Loss2	SinGAN-DNSR Loss3
	ns with field light rectang minimize the be located on	 15.20/0.75	 15.70/0.76	 15.80/0.77
	n or they ca h better fron e instruction is used. G hht rectangls	 12.68/0.72	 13.93/0.73	 14.50/0.78

Table 6.8 : Qualitative results for ablation experiment on different settings of SinGAN-DNSR for Kaggle Dirty document dataset along with their PSNR and SSIM values. Here, SinGAN-DNSR Loss1 corresponds to SinGAN-DNSR with original loss formulation, SinGAN-DNSR Loss2 corresponds to SinGAN-DNSR with perceptual loss function and SinGAN-DNSR Loss3 corresponds to SinGAN-DNSR with perceptual and style transfer loss function

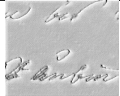
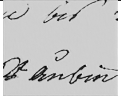
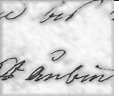
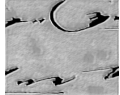
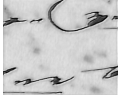
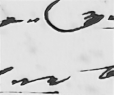
Input	Ground Truth	SinGAN-DNSR Loss1	SinGAN-DNSR Loss2	SinGAN-DNSR Loss3
	d bel It benfor	 10.76/0.50	 13.30/0.83	 15.78/0.87
	a a	 9.10/0.66	 9.68/0.79	 12.13/0.80

Table 6.9 : Qualitative results for ablation experiment on different settings of SinGAN-DNSR for DIBCO dataset along with their PSNR and SSIM values. Here, SinGAN-DNSR Loss1 corresponds to SinGAN-DNSR with original loss formulation, SinGAN-DNSR Loss2 corresponds to SinGAN-DNSR with perceptual loss function and SinGAN-DNSR Loss3 corresponds to SinGAN-DNSR with perceptual and style transfer loss function

এমন অপ্রত্যাশিত দৃশ্য দেখব, তা স্বপ্নেও ভা
দৃশ্য দেখলে একটুও অবাক হতাম না কিন্তু
অভাবনীয়, অচিন্তনীয়।

ছুটি পেলেই দিল্লি থেকে কলকাতায় ছুটে
আমার জন্ম, যেখানে আমার শৈশব-কৈশে
দিনগুলির অসংখ্য স্মৃতি ছড়িয়ে আছে, সে
এখানে এলেই পুরোনো বন্ধুদের কাছে
মেতে উঠি। সবাই নিলে সিনেমা-থিয়েটার
খাওয়া-দাওয়া করি। সবাই না হলেও কয়েকজন
কোথাও বেড়াতে যাই।



Bangla document reference image

Natural scene reference image

Table 6.10 : The two reference images taken to analyse the effect of reference image in our model

are quite different from the Kaggle and DIBCO images, they impart their pattern to the obtained SR image. We can see that some colored texture is added and there is distortion of the characters. We observe that using a reference image from similar dataset and same script is important to obtain desirable results.

6.5 RESULT AND DISCUSSION

We compare the proposed SinGAN-DNSR model with our baseline SinGAN model and the ZSSR model on the Kaggle Dirty Document Dataset and the DIBCO Dataset. Table 6.12 gives the performance of these models in terms of PSNR and SSIM metric. Tables 6.13 and 6.14 show qualitative results for images from the Kaggle dirty document and the DIBCO dataset respectively. Compared to ZSSR and SinGAN, it can be seen that our model super-resolves images along with denoising and achieves some reduction in noise.

We have also evaluated our model on the basis of OCR by using the Pytesseract-OCR software. OCR needs clean images for its optimal performance, hence we evaluate our model for OCR. For this evaluation we have used only images from the Kaggle Dirty Document image database as it contains printed text whereas the DIBCO dataset images contain handwritten characters which cannot be recognized by Pytesseract. We perform OCR on noisy LR images, clean ground-truth images, and the super resolved versions obtained from ZSSR, SinGAN and our proposed SinGAN-DNSR. It was found that our model performs the best in contrast to ZSSR and SinGAN. Table 6.15 depicts the qualitative results on text recognition. Table 6.16 gives the confidence scores and Character Error Rate obtained from Pytesseract for all the denoised Kaggle dirty document images obtained from ZSSR, SinGAN and our proposed SinGAN-DNSR. For the SinGAN-DNSR model text is recognized with higher confidence.

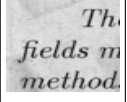
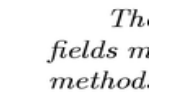
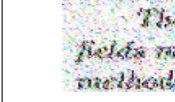
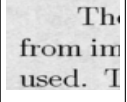
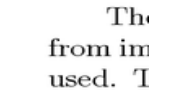
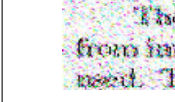
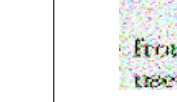
Input	Ground truth	Bangla Reference	Natural Reference
			
		12.66/0.35	12.67/0.37
			
		12.42/0.33	12.65/0.36

Table 6.11 : SR images obtained from SinGAN-DNSR model when a Bangla text document and a natural scene image are used as the reference image

Model Name	Kaggle Dirty	Document Dataset	BMVC Image	Quality Dataset
	PSNR	SSIM	PSNR	SSIM
ZSSR	15.86	0.67	10.45	0.67
SinGAN	16.23	0.66	10.47	0.68
SinGAN-DNSR	17.56	0.77	16.36	0.75

Table 6.12 : Comparison of results (PSNR and SSIM) on noisy low resolution document images on SinGAN, ZSSR and proposed SinGAN-DNSR

6.6 CONCLUSION

We have proposed a unified network for unsupervised SR and denoising of document images, which does not require a dataset of images to be trained on. The novelty of our model lies in the idea of freezing the discriminator trained on a different reference clean image, which would then be used to remove noise from the test image by contributing an additional noise penalty term in the loss function. The results of SinGAN-DNSR have shown consistent improvements over the results of vanilla SinGAN in terms of PSNR and SSIM. The applicability of our model is also shown in text recognition, where our model has resulted in higher confidence on recognized words. We have shown that the choice of a reference clean image is important in obtaining good denoising. Our results show a promising direction to further explore zero shot denoising and super-resolution.

In the next chapter, we address the problem of non-uniformly distributed noise component in the document image.

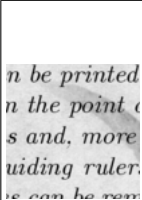
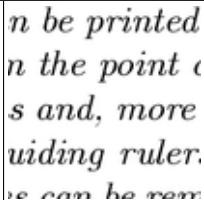
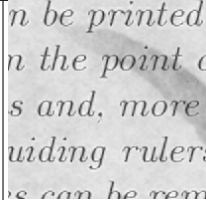
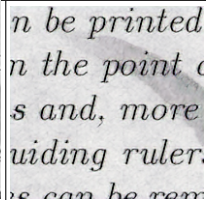
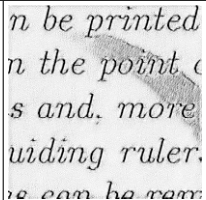
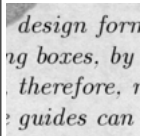
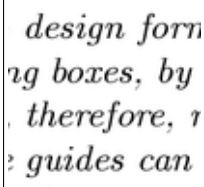
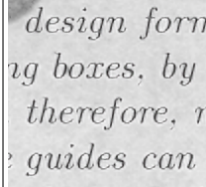
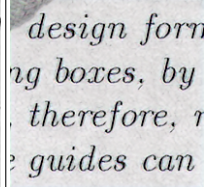
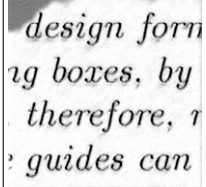
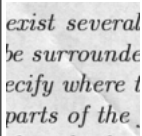
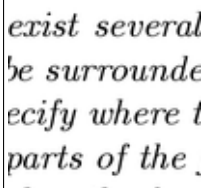
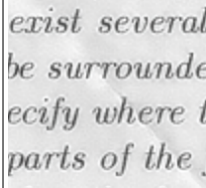
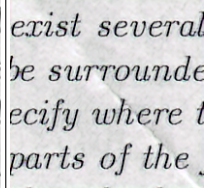
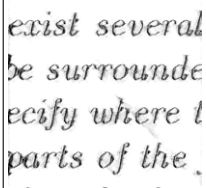
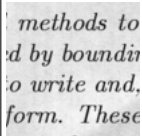
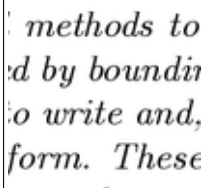
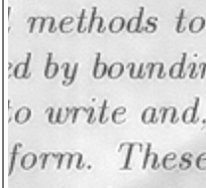
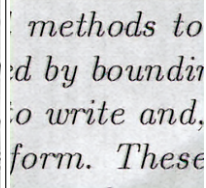
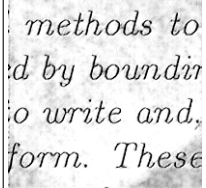
Input	Ground Truth	ZSSR	SinGAN	DNSR
				
		14.70/0.63	14.54/0.53	16.89/0.76
				
		15.10/0.65	15.87/0.73	17.02/0.76
				
		16.85/0.70	17.94/0.78	19.60/0.86
				
		16.82/0.71	14.92/0.55	18.39/0.79

Table 6.13 : Qualitative results for ZSSR, SinGAN and DNSR for Kaggle dirty document images with their PSNR/SSIM values

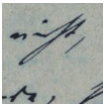
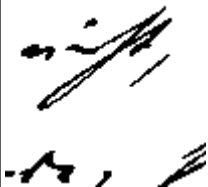
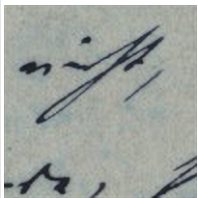
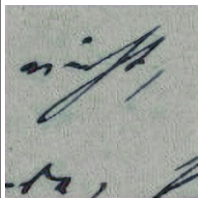
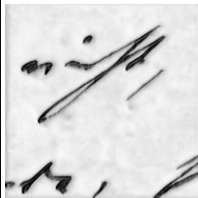
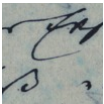
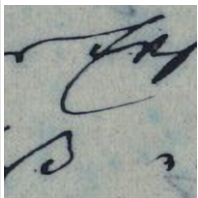
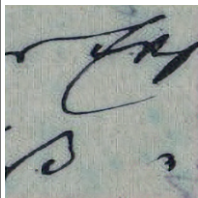
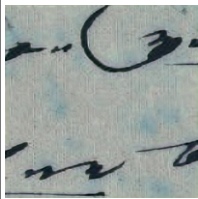
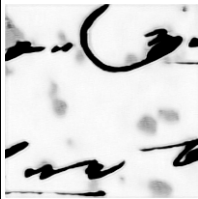
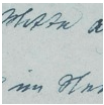
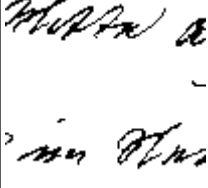
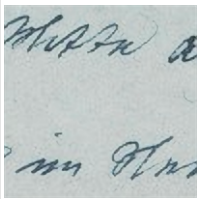
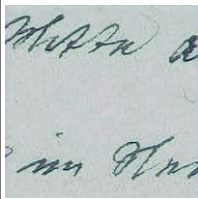
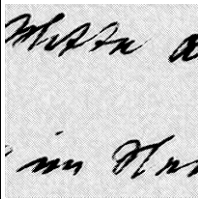
Input	Ground Truth	ZSSR	SinGAN	DNSR
				
		9.58/0.67	9.57/0.60	17.44/0.79
				
		9.53/0.66	9.57/0.59	15.82/0.71
				
		9.37/0.65	9.4/0.57	17.42/0.78
				
		13.35/0.77	13.33/0.74	14.76/0.72

Table 6.14 : Qualitative results for ZSSR, SinGAN and DNSR for DIBCO images with their PSNR/SSIM values

Input	Ground Truth	ZSSR	SinGAN	SinGAN-DNSR

Table 6.15 : OCR results for ZSSR, SinGAN and DNSR for Kaggle dirty document images with confidence value of each recognised word

Model Name	Tesseract OCR Confidence Score	Character Error Rate
Noisy LR image	80.15	0.51
Clean HR image	89.69	0.31
ZSSR	82.76	0.41
SinGAN	83.02	0.38
SinGAN-DNSR	87.26	0.36

Table 6.16 : Comparison of OCR results (Confidence Score and Character Error Rate) on high resolution document images obtained from SinGAN, ZSSR and proposed SinGAN-DNSR.