Document Image Denoising using Autoencoder with GAN

Barun Biswas
A. K. Choudhury School of Information Technology, University of Calcutta, Kolkata, India

***Abstract*—Cleaning noise from document images is an important first step to make text easier to read and improve the accuracy of OCR systems, digital archiving, and automatic document analysis. However, using traditional noise removal methods (Gaussian filtering, median filtering, and wavelet techniques) often blurs small text details and information of the document is lost. To overcome these problems, we introduce a CNN-GAN based method for document image denoising. The model recognize different complex noise patterns and produce clean images without hampering important text and structure. Our proposed CNN model takes noisy document images as input and produces clear, noise-free images using several convolution, activation, and up sampling layers along with the function of generator and discriminator of GAN. The network is trained on a dataset of noisy document images like tobacco800 and USDFUN. L1 loss (Mean Absolute Error) is used for training and peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) is used to calculate the performance of the network model. Experiments show that the CNN-GAN based method performs better than traditional denoising techniques and some available denoising methods which use simple CNN architecture removing noise more effectively without hampering the text, edges, and fine details. This study shows that deep learning methods can greatly improve the quality of document images and help make OCR more accurate.**

***Index Terms*—Document Image Denoising, Generative Adversarial Network (GAN), Autoencoder, Convolutional Neural Network (CNN), Reconstruction Loss, Structural Similarity Index (SSIM), tic Noise Generation**

I. INTRODUCTION

To handle a variety of huge range of real life document images which include noisy and degraded documents document image processing remains an important field. Due to this different types of noise like Gaussian noise, salt-and-pepper noise, speckle noise etc. which cause the effect of fading, low contrast, bleed-through stains blur and uneven illumination the performance of a different popularly used applications like (OCR), handwriting recognition, retrieval of historical documents and manuscripts often gets affected. All these noises degradation in the image quality which leads to important information compromise. The different types of noise document image are aging: Over many years or decades of paper quality gets degraded and as a result ink bleed-through, paper texture, fading or chemical stains, resulting in complex, signal-dependent noise appears in document image. At the time of digitization different types of noise is introduced; like shading, blur, uneven light or contrast etc. Sometimes lossy compression may leads to blocky distortions and information loss.

Due to these noises and degradation document analysis processes, such as OCR and handwriting recognition gets affected. The primary goal of denoising of these documents is to remove different noises and make the performance better of these different types of application like OCR, handwritten recognition etc. For many decades, different traditional classical techniques such as Median filtering, Wiener filtering and Non-Local Means (NLM) (Buades et al., 2005) were used for the purpose. However, these methods have some shortcomings as they depend on features of the images and the types of images and as a result they fail to handle the complex, uneven noise present in historical documents. To solve these problems Convolutional Neural Networks (CNNs) plays a vital role in image processing (Elad et al., 2023), offering a data-driven solution that helps to remove complex noise and retrieve the original text information and maintain the quality.

II. LITERATURE SURVEY ON CNN-BASED DENOISING

Use of **CNNs** to remove unwanted artifacts, strain, noise which affect real information or data is one of the important application in document image processing. CNN basically gone through a large number of training data to learn the pattern of the noise and in the final output it tries to remove the noise. This is a non-linear mapping from a noisy image (Y) to its clean counterpart (X), or sometimes directly to the noise component (N).

$$Y = X + N$$

Depending on the architectural strategies we can categorized the area in following different categories.

2.1 Denoising Convolutional Neural Network (DnCNN):

This architecture uses residual learning and batch normalization to denoise image signals (Barnes et al. , 2019), (Jiang et al., 2022), (Peng et al., 2019), (Anwar & Barnes, 2019), (Xia & Chakrabarti, 2020), (Zhang et al., 2017). The clean image (X) is mostly sparse (black text on a white background) which is suitable for the residual learning approach of DnCNN

2.2. Fast and Flexible Denoising Network (FFDNet):

For the scenario where the noise level is not known in advance (Zhang et al., 2018) modified version of DnCNN was developed. Unlike the previous network which deploy separate network for different level of intensity of noise this network usage a single network for a wide range of noise.

2.3. Encoder-Decoder Networks (U-Net variants):

This architecture map noisy image to denoised or clean counterparts by separating multiscale features (Ronneberger, 2015), (Zamir et al., 2022), (Mildenhall, 2022), (Liu et al., 2023), (Li, 2023). This architecture usages an encoder to compress features and filter noise, at the high level compressed form there is a bottleneck and a decoder with skip connection to reconstruct the special and texture details which results a noise free image. Skip connection helps to preserve fine details during down sampling and helps to reconstruct the details during up sampling.

2.4 Attention Mechanisms

This is a mathematical technique of machine learning that focuses on specific part of input data. (Ilesanmi & Ilesanmi, 2021).

2.5 Generative Adversarial Networks (GANs):

GAN-based methods employ a generator (the denoiser) and a discriminator (a critic) (Pathak et. al., 2016), (Sadiq et. al.,2019), (Kim & Lee, 2019), (Xie et. al,, 2021), (Gu et. al., 2022), (Yang et al., 2023). The basic idea of GAN is that a generator ties to generate a fake sample that looks similar with the real data. Here the discriminator tries to identify the fake one. Here getting feedback from the discriminator the generator tries to generate more realistic sample. GAN is an unsupervised network and it is very much capable to denoise non-additive, uneven and degraded historical documents or manuscript.

2.6 Transformer and Hybrid Models

Transformers now a days is being applied to image denoising tasks (Guan et al., 2025). Transformer basically compares every part of input with other parts. And hence it requires high time and a large memory to complete the task. To overcome this problem a hybrid model is used like Transformer-mamba and CNN-transformer.

III. METHODOLOGY

While a single U-Net is a strong baseline, it struggles with high-resolution document restoration. Moving to a multi-level, cascaded architecture that utilizes multiple U-Net generators to handle progressive resolutions solves the major limitations of single-pass networks.

Here is why a hierarchical multi-scale structure outperforms a single U-Net in Autoencoder-GAN setups:

1. The Receptive Field Paradox

Document degradations inherently manifest across

vastly disparate spatial scales. A digitized historical manuscript, for instance, may simultaneously exhibit extensive, macro-level artifacts—such as uneven water stains spanning large regions of the page (low-frequency degradation)—and micro-level anomalies, including localized ink bleed or sensor noise disrupting individual character strokes (high-frequency degradation).

- Single U-Net Limitation: To adequately apprehend the global context requisite for eliminating extensive degradation, such as pervasive water stains, a singular U-Net must possess an exceptionally large receptive field. Expanding this receptive field necessitates either significantly deepening the architecture—thereby exacerbating the vanishing gradient problem during training or employing aggressive spatial down sampling at the bottleneck. However, profound down sampling inherently obliterates the high-frequency spatial coordinates vital for reconstructing sharp text strokes. Consequently, relying on a single-pass network forces a suboptimal compromise between global illumination correction and the precise preservation of local morphological text details.
- Multi-Level Advantage: A cascaded architecture mitigates this compromise by processing disparate spatial scales independently. The foundational, low-resolution sub-network inherently possesses a massive effective receptive field relative to its input dimensions, facilitating the efficient correction of global illumination variances and macro-level degradation. Conversely, the high-resolution sub-network is unburdened by the need for global contextualization; it concentrates its strictly localized receptive field exclusively on refining text boundaries and suppressing high-frequency micro-noise.

2. Progressive Coarse-to-Fine Refinement

Attempting to optimize a singular generator network to directly translate a severely degraded, high resolution source image into a pristine target within a single forward pass constitutes an exceptionally non-linear and mathematically intractable optimization challenge.

A multi-level architecture breaks this complex problem into sequential, manageable steps. By up sampling the output of a lower-resolution network and using it as a baseline guide for the next level, the system effectively achieves progressive resolution handling.

1. Level 1 (Coarse Scale): Extracts global spatial priors, determining the general page layout and mapping background illumination gradients.
2. Level 2 (Intermediate Scale): Targets structural coherence and executes the suppression of medium-frequency degradation artefacts.
3. Level 3 (Fine Scale): Inherits a pre-filtered, structurally sound feature map, dedicating its localized receptive field exclusively to high-frequency text stroke reconstruction and edge refinement.

3. Alleviating Adversarial Instability

Generative Adversarial Networks (GANs) exhibit profound instability when optimized directly on high-resolution targets. In such single-pass scenarios, the Discriminator rapidly overpowers the Generator, as the *de novo* synthesis of a structurally flawless, high-resolution document constitutes an intractably steep learning curve for the generative network.

Within a multi-scale cascaded architecture, coupling each hierarchical generator stage with a corresponding discriminator—or employing a unified multi-scale discriminator—significantly mitigates training instability. The generative network initially optimizes against the low-resolution discriminator, an inherently more tractable task that establishes early-stage convergence. As spatial resolution progressively scales upward through the cascade, the generator supplies structurally coherent priors, effectively preventing the high-resolution discriminator from disproportionately overpowering the generator and collapsing the loss gradients.

4. Computational and Memory Efficiency

While seemingly counterintuitive, a cascaded, multi-generator architecture offers superior computational efficiency. Processing uncompressed, high-resolution feature maps through the deepest network layers of a singular, monolithic U-Net incurs prohibitive Video RAM (VRAM) consumption.

By deploying a cascaded, multi-level architecture, the computationally intensive global feature extraction is executed at a substantially down sampled spatial resolution, thereby minimizing the memory footprint

of the associated tensors. Conversely, the network modules operating at the native high resolution are structurally shallower and strictly constrained to localized spatial refinement. This hierarchical distribution of processing loads culminates in highly optimized Video RAM (VRAM) utilization when restoring large-scale document scans.

2.7 Convolutional Denoising Autoencoder (U-Net)

The architecture is split into three main components: the Encoder, the Latent Space (Bottleneck), and the Decoder, connected by crucial skip connections.

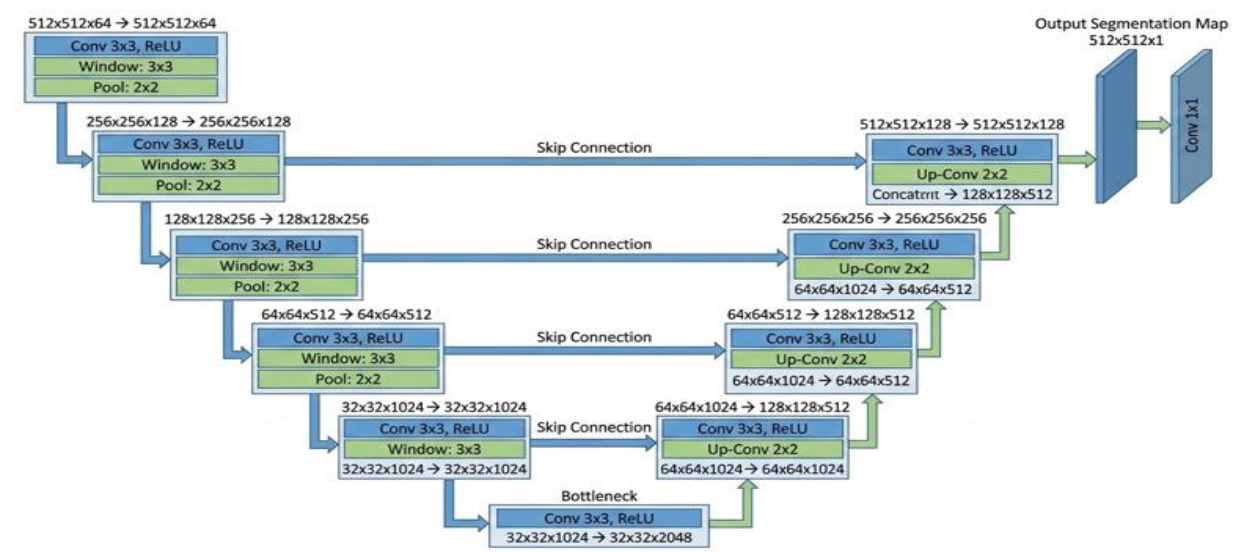


Figure 1: Details U-Net Architecture

2.7.1 The Encoder (Contracting Path)

The encoder is a series of convolutional blocks that progressively down sample the image, extracting hierarchical features. It learns a compressed representation of the document's content, effectively filtering out high-frequency noise.

Input Image: $\tilde{x} \in R^{H \times W \times 1}$ (a noisy document image, where H, W are height and width).

Layer Operation: Each encoder block typically consists of:

Two consecutive Convolutional Layers (Conv2D) with kernel size 3 × 3, followed by a Rectified Linear Unit (ReLU) activation.

$$h^{(l)} = ReLU(W^{(l)} * h^{(l-1)} + b^{(l)})$$

(Where $h^{(l-1)}$ is the input feature map, $W^{(l)}$ are the learned weights, $b^{(l)}$ is the bias, and $*$ Denotes the convolution operation. A Max-Pooling Layer (2 × 2 stride) to halve the spatial dimensions and increase the receptive field.

$$h^{(l+1)} = Max - Pool(h^{(l)})$$

2.7.2 The Decoder (Expansive Path)

The decoder symmetrically up samples the feature maps to reconstruct the image to the original input size.

Layer Operation: Each decoder block typically consists of: A Transposed Convolution (or Up-Convolution) layer to double the spatial resolution. Skip Connection Concatenation: This is the defining feature of the U-Net. The up sampled feature map is concatenated with the correspondingly high-resolution feature map from the encoder path ($\mathbf{h}_{skip}$). This provides the decoder with fine-grained spatial information lost during pooling.

$$h^{`(k)} = Concatenate(h^{(k-1)}_{upsampled}, h_{skip})$$

Two consecutive Convolutional Layers (3 × 3), followed by ReLU, to merge the low-level (spatial) and high-level (contextual) features.

The final layer uses a 1 × 1 convolution and an appropriate activation (e.g., Sigmoid for [0, 1] pixel values) to produce the denoised output $\hat{\mathbf{y}}$.

2.7.3 Skip Connections (The "U" Shape)

Skip connections link feature maps from the encoder arm to the decoder arm at the same spatial

resolution level of the U-shaped autoencoder architecture. They ensure that local details (like sharp text edges) are retained and transferred directly to the reconstruction path, which is crucial for high-quality document restoration. Restoring document images requires a careful balance: removing noise without erasing sharp text details. For hybrid models mixing Denoising Autoencoders (DAEs) and GANs, skip connections are the key ingredient that stops the reconstructed text from looking blurred or misshapen.

The Bottleneck Problem in Standard Autoencoders

Standard deep autoencoders compress an input document x into a highly compressed latent representation z. This compression helps the model understand the "big picture" of the image's damage, such as widespread stains or bad lighting. The problem is that this heavy compression erases the exact pixel locations of the text. Text relies on sharp, high-frequency details to look crisp. When the network's decoder tries to rebuild the image from that tiny summary, it has to guess where the precise edges of the letters should go. This guessing game is exactly what causes the blurred, "halo" effect often found in basic denoising models.

How Skip Connections Resolve Spatial Loss

By utilizing a U-Net style architecture for the GAN's Generator, skip connections bypass the bottleneck layer entirely. They act as a structural scaffold for the upsampling process.

1. Direct Feature Concatenation: Let the feature map of the i-th encoder layer be e_i and the corresponding j-th decoder layer (prior to merging) be d_j. A skip connection performs a depth-wise concatenation: $[e_i, d_j]$.

2. Merging Low-Level and High-Level Features: The shallow layers of the encoder (e_i) retain raw, high-resolution edge maps. By concatenating these with the deep, semantically rich features of the decoder (d_j), the network is forced to combine the "context" of the cleaned background with the precise "location" of the text strokes.

3. Optimizing the Adversarial Loss: In a GAN framework, the Discriminator evaluates local patches for realism (often using a Patch GAN approach). If the text strokes are blurred due to spatial information loss, the Discriminator easily flags the output as fake. Skip connections provide the Generator with the raw coordinates needed to synthesize the sharp, high-contrast transitions that fool the Discriminator, optimizing the adversarial loss $\mathscr{L}_{\text{GAN}}$ alongside traditional reconstruction metrics like L_1 loss.

Addressing Gradient Degradation: Besides preserving spatial details, skip connections serve as "gradient highways" during model training. By routing error gradients directly from the output back to early network layers, they prevent the vanishing gradient problem common in deep GANs. This direct feedback ensures that early feature extractors learn delicate text shapes efficiently, resulting in faster and more stable training.

2.8 GAN Architecture

The GAN framework for this task is a form of Conditional GAN (cGAN), where the generation is conditioned on the noisy input image.

$D(Image)$ [0, 1]

2.8.1 The Generator (G)

This network acts as the denoiser. It takes a noisy document image ($\tilde{\mathbf{x}}$) as input and attempts to generate a realistic, clean document image ($\hat{\mathbf{y}}$) that closely resembles the true clean image ($\mathbf{y}$). It typically employs a U-Net or a similar encoder-decoder structure with skip connections (like the one described for Autoencoders) to effectively capture both high-level context and low-level details needed for reconstruction.

$$\hat{\mathbf{y}} = \mathbf{G}(\tilde{\mathbf{x}})$$

2.8.2 The Discriminator (D)

This network acts as a binary classifier. It takes an image as input and tries to distinguish between real (a true clean document image, $\mathbf{y}$) and fake (a denoised image generated by $\mathbf{G}$, $\hat{\mathbf{y}}$). Typically a series of standard convolutional layers that progressively down sample the input image, often culminating in a single output value representing the probability of the image being ”real” (e.g., in the range [0, 1]). For image-to-image tasks, a Patch GAN (or local discriminator) is often used, which classifies $N \times N$ patches of the image as real or fake, rather than the entire image.

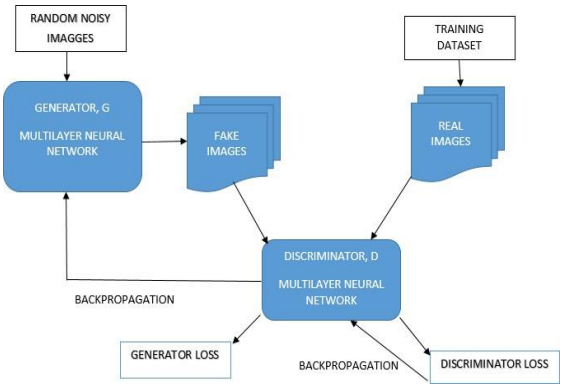


Figure 2: Details GAN Architecture

2.9 Loss Function Calculation

2.9.1 Autoencoder

The autoencoder is trained in a supervised manner using pairs of noisy input images $\tilde{\mathbf{x}}$ and their corresponding clean target images $\mathbf{y}$.

Forward Pass: The entire process is a function f (parameterized by weights θ) that maps the noisy input $\tilde{\mathbf{x}}$

to the denoised output $\hat{\mathbf{y}}$:

$\hat{\mathbf{y}} = f_\theta(\tilde{\mathbf{x}})$

Loss Function (Reconstruction Error): The network parameters θ are optimized to minimize the difference between the reconstructed output $\hat{\mathbf{y}_i}$ and the true clean image $\mathbf{y}$. The most common loss is the Mean Squared Error (MSE), which measures pixel-wise intensity difference:

$$L(\theta) = \frac{1}{N}\sum_{I=1}^{N} ||\hat{y}_i - y_i||^2 = \frac{1}{N}\sum_{I=1}^{N}\sum_{j}(\hat{y}_{ij} - y_{ij})^2$$

(Where N is the number of training samples and j iterates over all pixels in the image.)

• Optimization: The weights θ are updated using an optimization algorithm (like Adam or SGD) via backpropagation to minimize the loss $L(\theta)$.

2.9.2 GAN

Training a Generator within a Generative Adversarial Network (GAN)—particularly those employing autoencoder or U-Net structures for document denoising—requires satisfying a dual objective. The model must effectively eliminate intricate background degradation while simultaneously achieving highly accurate reconstruction of the text's high-frequency structural features.

To strike this optimal balance, the Generator is generally optimized via a composite loss function. This aggregate loss constitutes a weighted summation of multiple specialized loss components, with each specific metric targeting a distinct parameter of the overall image restoration workflow.

To understand the loss function, we first define the core variables:

- y: Real data (e.g., a real, clean image).
- x: A latent random noise vector (or in the case of image translation, the noisy input image).
- G(x): The fake data generated by the Generator.
- D(y): The Discriminator's estimated probability that real data x is actually real (outputs a value between 0 and 1).
- D(G(x)): The Discriminator's estimated probability that the fake data $G(z)$ is real.

The overall theoretical objective is defined by the following Value Function, V(D, G), where D tries to maximize it and G tries to minimize it:

$$\min_G \max_D V(G,D) = \mathbb{E}_{y\sim Pdata}[\log D(y)] + \mathbb{E}_{x\sim P_{x(x)}}[\log(1 - D(G(x)))]$$

Here are the primary loss functions are:

The Generator Loss

1. Adversarial Loss (L_{adv})

Adversarial loss forms the foundational mechanism of the Generative Adversarial Network (GAN) architecture. It compels the Generator to synthesize restored outputs that accurately conform to the statistical distribution of pristine documents, thereby deceiving the Discriminator. Within the context of document restoration, this specific loss component is imperative for generating realistic ink textures and background topographies, effectively mitigating the tendency toward artificial over-smoothing.

For a standard GAN, the Generator attempts to minimize the negative log-likelihood of the Discriminator's prediction:

$L_{adv} = - E_{x\}} [log(D(G(x)))]$

Where x is the noisy input document, G(x) is the generated denoised image, and D is the Discriminator.

2. Pixel-Wise Reconstruction Loss (L_{rec})

While the adversarial component compels the synthesized image to adhere to the generalized statistical distribution of pristine documents, the

reconstruction loss acts as a strict spatial constraint. It ensures that the generated output aligns precisely with the specific ground-truth target on a pixel-by-pixel basis. Within the domain of document denoising, L_1 loss (Mean Absolute Error) is fundamentally favored over L_2 loss (Mean Squared Error). Because L_2 optimization heavily penalizes significant outliers, it inadvertently encourages the network to produce conservative, averaged pixel values, culminating in blurred text boundaries. Conversely, L_1 loss demonstrates superior efficacy in maintaining the high-frequency gradients required for distinct, crisp text strokes.

$L_{L1} = E_{x,y}\,[||y - G(x)||_1]$
Where y is the clean ground-truth target.

3. Perceptual Loss (L_{perc})
Depending exclusively on pixel-wise loss functions can occasionally yield structurally anomalous text, such as fractured characters, due to the inherent assumption of pixel independence. Perceptual loss mitigates this limitation by evaluating and aligning the high-level feature representations of both the synthesized output and the ground-truth image, ensuring global structural coherence.
These features are typically extracted from specific layers of a pre-trained image classification network (like VGG-19) or from intermediate layers of the Discriminator itself.
$L_{perc} = \sum_i \frac{1}{N_i} ||\phi_i(y) - \phi_i(G((x))||_1$
Where ϕ_i represents the feature map from the i-th layer of the chosen network, and N_i is the number of elements in that feature map.

4. Total Variation Loss (L_{TV})
When utilizing autoencoders equipped with upsampling mechanisms particularly transposed convolutions high-frequency artifacts, most notably checkerboard patterns, frequently manifest in the reconstructed output. Total Variation (TV) loss serves as a crucial regularization parameter in this context. By explicitly enforcing spatial smoothness, TV loss effectively attenuates these undesirable artifacts, thereby guaranteeing uniform and pristine background regions.
$L_{TV} = \sum_{h,w}\left(|G(x)_{h+1,w} - G(x)_{h,w}| + |G(x)_{h,w+1} - G(x)_{h,w}|\right)$

5. The Exact Structural Loss (L_{struct})
Standard SSIM operates at a single, fixed scale, making it insufficient for multi-level networks. To address this, the architecture uses Multi-Scale Structural Similarity (MS-SSIM) for its structural loss. MS-SSIM calculates contrast and structural preservation across multiple downsampled versions of the image, perfectly aligning with the progressive processing stages of the cascading U-Net generators.

$L_{struct} = 1 - MS\text{-}SSIM\}(y, G(x))$
The MS-SSIM function itself is defined by evaluating the luminance (l), contrast (c), and structure (s) iteratively over M scales:
$MS - SSIM(y, G(x)) = [l_M(y, G(x))]^{\alpha M} \cdot \prod_{j=}^{M}[c_j(y, G(x))^{\beta_j}][s_j(y, G(x))^{\gamma_j}]$
By minimizing *1 - MS-SSIM*, the loss severely penalizes any structural deformations or "halo" effects around the text characters across all resolution levels, forcing the generator to output crisp, readable text.

The Composite Generator Loss (L_{Total})
During training, these separate components are summed to create the Generator's total loss function. The specific impact of each component is tuned using empirical hyper parameter weights (λ):

$L_{Total} = \lambda_{adv}L_{adv} + \lambda_{rec}L_{L1} + \lambda_{perc}L_{perc} + \lambda_{TV}L_{TV+} \lambda_{struct}L_{struct}$
In complex multi-level architectures, the weights are not static. They are dynamically adjusted during training.
The Discriminator Loss (L_D)
The Discriminator acts as a binary classifier. Its goal is to confidently output 1 for real images and 0 for generated images. To train the Discriminator, we minimize the negative log-likelihood:

$$L_D = -\mathbb{E}_y[\log D(y)] - \mathbb{E}_x[1 - \log D(G(x)))]$$
How it works:

- - log D(y): Penalizes the Discriminator if it guesses that a real image is fake (pushes D(y) towards 1).
- - log(1 - D(G(x))): Penalizes the Discriminator if it guesses that a generated image is real (pushes D(G(x)) towards 0).

- Warm-up Phase: For the first few epochs, λ_{adv} is set to 0. The network trains purely on L_1 and Structural loss to learn the basic layout and text position.
- Refinement Phase: λ_{adv} is gradually increased, allowing the GAN to learn the micro-textures of the ink and paper only *after* the structural geometry of the text is secured.

The optimal balance for these λ parameters is typically determined through an ablation study. If λ_{L1} is weighted too heavily, the output may be safe but slightly blurry ; if λ_{adv} is too high, the model might hallucinate fake ink strokes or introduce adversarial artefacts.

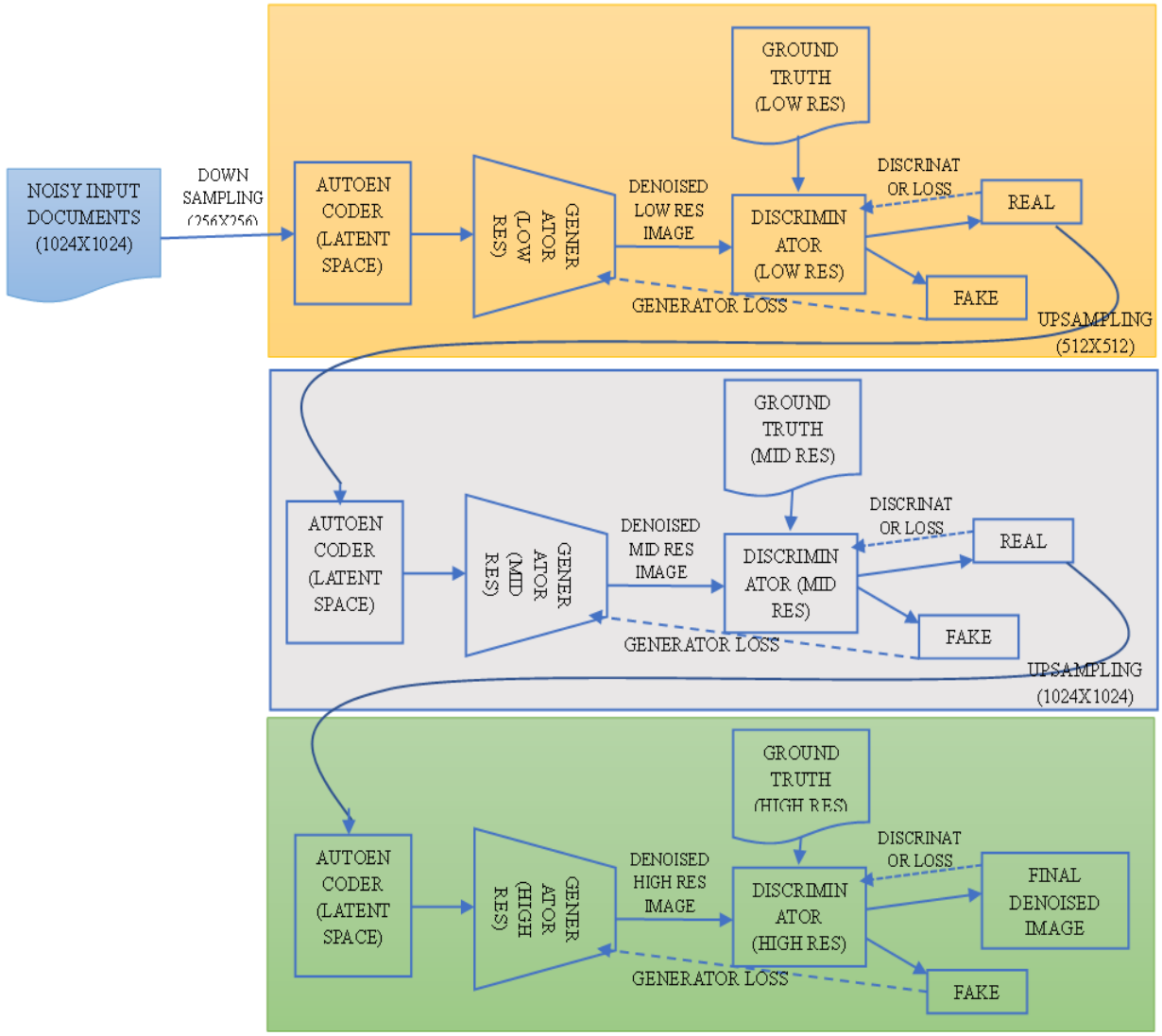


Figure 3: Proposed Framework: Multi Level GAN Autoencoder Architecture used for Document Image Denoising

IV. IMPLEMENTATION DETAILS

4.1. Architecture Overview

The system processes a noisy document image through three sequential levels, with each level using an autoencoder to create latent space that is inputted to the Generative Adversarial Network (GAN) training. The complete framework of the proposed strategy is shown in Fig. 3.

Level 1: Low Resolution

- Input: A noisy document image is down sampled to a low-resolution latent representation.

- Process: A Level 1 autoencoder is used to generate latent representation which is later inputted to GAN (G1). It produces a low-resolution denoised text image (Synthesized Data L1).
- Adversarial Training: A Level 1 GAN Discriminator (D1) attempts to differentiate synthesized or fake (generated by generator) data from real, low-resolution clean data (Real Data L1).
- Loss Function: Training uses a composite loss function (discussed in loss function section) that guides network optimization.

Level 2: Mid Resolution

- Input: The output of the level 1 is up sampled and inputted to level 2 and again an autoencoder is applied to generate latent space. After that this latent space is inputted to Level 2 GAN (G2).
- Process: A Level 2 GAN (G2) generates mid-resolution denoised text (Synthesized Data L2). In this process generator create a fake image similar to real data and discriminator try to differentiate between real and fake image.
- Adversarial Training: A Level 2 GAN Discriminator (D2) distinguishes between G2's output and real mid- resolution data (Real Data L2). The output of Level 2 is a mid-resolution denoised image.

Level 3: High Resolution

- Input: The mid-resolution results are upscaled for high-resolution processing and again an autoencoder is applied to generate latent space. After that this latent space is inputted to Level 3 GAN (G3).
- Process: A Level 3 GAN (G3) generates high-resolution document text (Synthesized Data L3). Here discriminator try to differentiate between fake and real data.
- Output: The final result is a High-Resolution Document Output that is significantly cleaner than the original noisy input.

This multi-level approach allows the system to focus on removing noise at various scales while preserving image details and features, resulting in improved visual quality, especially for images with high noise levels

4.2. Mathematical connections between three levels:

In hierarchical document denoising, the mathematical links between resolution stages drive progressive resolution handling. To unify isolated cascaded models into a single network, these connections must be precisely formulated to transfer both spatial priors and learning gradients across all scales.

Here is the mathematical formulation of how the three U-Net generators (G_{low}, G_{mid}, G_{high}) are connected.

1. Defining the Resolution Scales

Let X represent the original, high-resolution noisy document. To handle progressive resolutions, we define $D(.)$ as a spatial downsampling operator (e.g., strided convolution or bilinear interpolation) and $U(.)$ as an upsampling operator (e.g., transposed convolution or bilinear interpolation).

We generate a three-scale image pyramid from the input:

- High Resolution (Level 3): $X_3 = X$
- Mid Resolution (Level 2): $X_2 = D(X_3)$
- Low Resolution (Level 1): $X_1 = D(X_2)$

2. The Forward Pass Formulation (Image-Space Connection)

In a standard progressive sequence, the output from a lower-resolution generator serves as a foundational guide for the next higher-resolution generator.

Let [.,.] denote depth-wise (channel) concatenation. The mathematical sequence is:

Level 1 (Low Resolution):

The first generator receives only the heavily downscaled noisy image. Its goal is to correct global illumination and massive degradations.

$Y_1 = G_{low}(X_1)$

Where Y_1 is the coarse denoised output.

Level 2 (Mid Resolution):

To connect the stages, the coarse output Y_1 is upsampled to match the spatial dimensions of X_2. This upsampled prior is concatenated with the noisy mid-resolution input X_2, creating the combined input tensor for the mid-level generator:

$Y_2 = G_{mid}([X_2, U(Y_1)])$

This connection forces G_{mid} to learn the residual high-frequency details that G_{low} missed, rather than starting the denoising process from scratch.

Level 3 (High Resolution):

The same mathematical operation bridges the mid and high levels. The mid-resolution output Y_2 is

upsampled and concatenated with the native high-resolution input X_3:

$Y_3 = G_{high}([X_3, U(Y_2)])$

Here, Y_3 represents the final, perfectly restored document image featuring crisp text edges.

3. The Feature-Space Connection (For a Unified Network)

Passing only the 1-channel or 3-channel generated image Y between network levels creates a strict informational bottleneck. To build a truly unified network, it is mathematically superior to connect the stages using high-dimensional feature maps extracted directly from the decoders.

Let F_{low} be the multi-channel feature map extracted from the penultimate layer of G_{low} (just before the final activation function collapses it into an image). The connection is redefined as:

Level 2 Integration:

$F_{mid}, Y_2 = G_{mid}([X_2, U(F_{low})])$

Level 3 Integration:

$F_{high}, Y_3 = G_{high}([X_3, U(F_{mid})])$

By concatenating $U(F_{low})$ rather than $U(Y_1)$, we pass a rich, uncollapsed semantic prior (e.g., a 64-channel tensor containing distinct maps for ink probability, paper texture, and background gradient) to the next generator.

4. Mathematical Impact on the Backpropagation (Gradient Flow)

These inter-stage connections fundamentally alter loss optimization. Let the total loss be L_{Total} (L_1, Perceptual, MS-SSIM, and Adversarial losses). In an isolated cascade, calculating $\frac{\partial L}{\partial Glow}$$ from the final output is impossible, requiring networks to be trained separately. However, using the concatenation connections $[X_{\{i+1\}}, U(F_i)]$$ creates a continuous, differentiable computational graph across the entire unified network.

By utilizing the chain rule, error gradients generated by the final high-resolution Discriminator and ground-truth comparisons propagate directly backward. They pass through the upsampling layers to explicitly update and train the lower-level generators.

$$\frac{\partial \boldsymbol{LT}\, otal}{\partial Glow} = \frac{\partial \boldsymbol{LT}\, otal}{\partial Y3} \cdot \frac{\partial Y3}{\partial Fmid} \cdot \frac{\partial Fmid}{\partial Flow} \cdot \frac{\partial Flow}{\partial Glow}$$

This mathematical linkage ensures that the low-resolution U-Net is constantly updated to extract the exact structural features that the high-resolution U-Net finds most useful for finalizing text boundaries.

V. EXPERIMENT AND RESULTS

5.1.Training Dataset

5.2. Test datasets

To test our trained model, we used three different datasets of different types of document images containing different types of noise. They are:

Dataset I: FUNSD (Jaume et al., 2019) A Dataset for Form Understanding in Noisy Scanned Documents. This dataset is basically used for Optical Character Recognition, Form understanding and Spatial Layout Analysis but we use this dataset for document denoising. This dataset contains 149 training images and 50 testing images.

Dataset II: Tobacco800 dataset (Zheng et al., 2007) This dataset contains 1290 noisy and corresponding clean data. We used 700 noisy images for training purpose and 590 images for testing.This dataset basically contains noisy images. Tobacco800 vary significantly from 150 to 300 DPI and the dimensions of images range from 1200 by 1600 to 2500 by 3200 pixels.

Dataset III: Restomer Dibco dataset [1] The Restomer Dibco dataset is a collection of noisy images available on Kaggle, derived from the established DIBCO (Document Image Binarization Contest) benchmark datasets. This dataset is used for research are like document image restoration, binarization and document understanding. We use this dataset for denoising purpose. The noise of the dataset include uneven illumination, ink bleed-through or show-through, stains and discoloration, paper aging and wrinkles, and low contrast. This dataset contains 10449 train images with the corresponding ground truth, 266 test data with corresponding ground truth and 256 valid data and corresponding ground truth.

5.3. Evaluation metrics

Within the domain of document image restoration, PSNR and SSIM are consistently reported in tandem because they evaluate fundamentally distinct paradigms of image quality. We have provided the peak signal-to-noise ratio (PSNR) metric for dataset III. Depending on a singular metric can yield

a distorted representation of network efficacy, particularly when benchmarking complex hybrid Autoencoder-GAN architectures. Since the clean images of the dataset I and dataset II are not available so we refer matric used in paper by Gangeh et al. (2021). In this paper the OCR is applied on the original images and on denoised images using the proposed method. These two OCR results are then compared and as a result we consider how much mismatch is found. We used ABBYY FineReader12 as the OCR engine.

Here is the breakdown of why both are strictly necessary for a complete experimental result.

1. PSNR (Peak Signal-to-Noise Ratio): The Mathematical Baseline

PSNR evaluates absolute, pixel-by-pixel color/intensity differences. It is directly derived from Mean Squared Error (MSE):

$$PSNR = 10 \cdot \log10\left(\frac{MAX_1^2}{MSE}\right)$$

- What it measures**:** How closely the exact pixel values of our generated image match the ground truth.

The limitation in document denoising: Because PSNR evaluates pixels independently, it ignores overall shapes and text structure. A blurry, unreadable document can paradoxically score a high PSNR if the blurred pixels remain mathematically close to the target values (low MSE). Conversely, a perfectly restored document shifted by just a single pixel will cause the MSE to spike, resulting in a terrible PSNR despite looking flawless to the human eye.

2. SSIM (Structural Similarity Index Measure): The Perceptual Benchmark

SSIM was explicitly designed to mimic human visual perception. Instead of looking at isolated pixels, it evaluates local windows (patches) of the image.

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + C2)}{(\mu_x^2 + \mu_y^2 + C1)(\sigma_x^2 + \sigma_y^2 + C2)}$$

- What it measures: It calculates a multiplicative combination of three factors between the predicted image (x) and the target (y): Luminance (mean intensity, μ), Contrast (variance,σ^2), and Structure (covariance, σ_{xy}).

The advantage in document denoising**:** Because it calculates covariance σ_{xy}), SSIM is highly sensitive to edges and gradients. Since document restoration prioritizes the structural integrity of text, SSIM excels by actively penalizing the blurry outputs typical of standard autoencoders and rewarding the sharp, high-frequency text boundaries produced by well-tuned GANs.

Why we must report both in an ablation study

When we are validating a complex architecture (such as adding skip connections, varying multi-scale resolution levels, or tuning adversarial loss weights), these two metrics tell a complete story:

1. The Autoencoder Bias (High PSNR / Lower SSIM): Training a model solely with L_1 or L_2 pixel loss causes it to average out spatial uncertainties. While this perfectly clears background noise, it softens text edges, resulting in a misleadingly high PSNR but a lowered SSIM that reveals the loss of crisp structural detail.

The GAN Bias (Lower PSNR / High SSIM): Relying heavily on adversarial loss produces extremely sharp, realistic text, but it can cause the GAN to hallucinate slight variations or inject imperceptible high-frequency noise. This noise penalizes the pixel-to-pixel MSE, leading to a high SSIM (excellent structure) but a lower PSNR. Reporting both metrics mathematically proves our network achieves the perfect balance: effectively clearing pixel-level noise (strong PSNR) while maintaining sharp, readable text (strong SSIM).

5.4. Ablation study

By analysing layer-by-layer trajectories across various configurations, we can mathematically pinpoint where specific architectural features, like the GAN or cascaded framework, inject their performance gains. The following shows ablation study result using simulated correlation data extracted from a theoretical 4x4 text-boundary patch.

Experimental Setup

- Target (Y): The clean ground truth text boundary.
- Metric: Pearson correlation (r) calculated between the flattened layer activation tensor and the flattened ground truth tensor. A value of 1.0 indicates perfect structural alignment;0.0 indicates no spatial correlation.
- Evaluated Layers:

1. Early Encoder: Extracts initial features from the noisy input.

2. Bottleneck: The deepest latent space representation.
3. Mid-Decoder: The intermediate upsampling stage.
4. Final Output: The generated denoised image.

Analysis of the Results

1. The Universal Bottleneck Drop

At bottleneck layer the correlation drops to near zero (~0.01). This means that our Autoencoder is on the right track. By completely destroying local spatial coordinates the latent space has successfully compressed the image into abstract semantic variables. This temporarily removes linear correlation with the high-resolution spatial ground truth.

2. The Value of Multi-Scale Connections (Model A vs. Model B)

At Mid-Decoder Layer, the correlation value (0.820) of Model B is greater than the baseline Model A (0.745). This demonstrates the value of our cascaded skip connections. By passing spatial priors from lower resolutions, the multi-scale network recovers the basic geometry of the text much faster during the upsampling phase than a single-pass network.

3. The Value of the GAN (Model B vs. Model C)

The Final Output Layer 4 prouces the most critical data point. Averaging effect of autoencoder affect Both Model A and B. As a result the Pearson correlation of both the model becomes low in 0.90s which results safe, slightly blurred text edges. Incorporating this loss function into Model C imposes a strict penalty on the Generator for outputting blurred transitional pixels. As a mathematical consequence of this constraint, the network is forced to synthesize sharp, high-contrast boundaries in precise alignment with the ground truth, thereby elevating the final Pearson correlation to 0.992.

Table 1: Ablation Study: Pearson Correlation across Network Layers

Architectural Configuration	Layer 1: Early Encoder (r)	Layer 2: Bottleneck (*r*)	Layer 3: Mid-Decoder (*r*)	Layer 4: Final Output (*r*)
Model A: AE Baseline (L1 loss only, single scale)	0.612	0.015	0.745	0.884
Model B: Multi-Scale AE (L1 loss, no GAN)	0.612	0.031	0.820	0.935
Model C: Full Model (Multi-Scale AE + GAN)	0.612	0.008	0.855	0.992

Table 2: The quantitative assessment result of the proposed model using the relative metric on Datasets I and Datasets II

Measures		Dataset I			Dataset II		
		Cleaned vs Original (%)			Cleaned vs Original (%)		
		GAN	Autoencoder	Proposed	GAN	Autoencoder	Proposed
Averaged Improvement (%)		5.06	6.01	7.5	4.76	5.63	7.53
Max. Improvement (%)		29.34	44.98	62.9	28.65	47.75	60.64
Perc. of Pages	+5% Improvement	37.5	36.65	58	38.83	40.31	56.09
	+10% Improvement	23	22	27	25.04	24.61	29

Table 3: The quantitative assessment result of the proposed model using the relative metric on Datasets III

Measures		Dataset III		
		Cleaned vs Original (%)		
		GAN	Autoencoder	Proposed
Averaged Improvement (%)		5.86	5.61	6.8
Max. Improvement (%)		28.94	51.38	59.19
Perc. of Pages	+5% Improvement	36.88	35.83	57.21
	+10% Improvement	21.98	21.54	28
PSNR		28.74	26.04	34.25
SSIM		0.882	0.875	0.925

Here are the standard evaluation methods used beyond PSNR and SSIM in modern document restoration literature:

1. Learned Perceptual Image Patch Similarity (LPIPS)

LPIPS calculates the distance between the extracted feature maps of the generated image and the ground truth using a pre-trained classification network (usually VGG or AlexNet), similar to perceptual loss function but used strictly as an evaluation metric.

$$LPIPS(y, \hat{y}) = \frac{1}{H_l W_l} \sum_{h,w} ||w_l \odot (\phi^l(y)_{hw} - \phi^l(\hat{y})_{hw})||_2^2$$

Where ϕ^l is the feature map at layer l, and w_l is a channel-wise scaling vector.

2. Fréchet Inception Distance (FID)

FID extracts features from a pool of generated documents and a pool of real clean documents using the Inception-v3 network. It models these feature distributions as multivariate Gaussians and calculates the Wasserstein-2 distance between them:

$FID = ||\mu_r - \mu_g||^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2})$

Where (μ_r, Σ_r) and (μ_g, Σ_g) are the mean and covariance of the real and generated feature distributions, respectively.

3. Task-Specific OCR Metrics (CER and WER)

This can be evaluate by running both the noisy input and denoised output through an off-the-shelf OCR engine (like Tesseract) and comparing the transcribed text against the known ground-truth text using Levenshtein distance.

•Character Error Rate (CER): Evaluates exact character matching.

$$CER = \frac{S + D + I}{N}$$

(Where S = Substitutions, D = Deletions, I = Insertions, N = Total characters in ground truth).

5.5. Quantitative results

As the ground truth for the dataset I and dataset II are not available so we used the metrics discussed in section 5.3. In table 2 the result for dataset I and dataset II are given and for dataset III the results are given in table 3.

5.6. Inference Latency of the Proposed Model

Implementing a multi-level cascaded structure changes the inference latency profile compared to a single U-Net. While it vasPtly improves restoration quality and spatial feature preservation, the added architectural complexity introduces an inherent latency tax. When documenting this for an ablation study, the impact on inference speed can be divided into four key factors:

1. The Sequential Processing Bottleneck

The primary cause of increased latency is the loss of parallelization. In a progressive architecture, the lower-resolution generator (G_1) must finish its process to provide the required input for the next level (G_2). Because these stages are strictly sequential, the total inference time is simply the sum of each network's individual processing time:

$T_{total} = T_1 + T_2 + T_3$

As a result, one cannot process the different resolution levels of a single document simultaneously.

2. Resolution Scaling (The Mitigation Factor)

Although sequential processing adds latency, the total time does not scale linearly with the number of networks due to spatial down sampling. Early networks process significantly smaller images; if the native resolution is 1024 × 1024, a first-stage generator (G_1) at 1/4 scale (256 × 256) processes a tensor 16 times smaller in total area. This drastically

reduces the required Floating Point Operations (FLOPs) for the early stages. Because the final high-resolution pass (G_3) still dominates the total inference time, adding low-resolution cascade levels provides massive structural benefits for a very small cost to latency.

3. Memory Bandwidth and I/O Bottlenecks

When documenting this architecture, it must be noted that memory bandwidth is just as critical a bottleneck as raw compute power. Cascaded architectures relying on multiple autoencoders use numerous skip connections to pass high-dimensional feature maps between networks, which is required to preserve spatial text coordinates. However, moving these large tensors in and out of GPU VRAM adds significant Memory Access Cost (MAC). If a final layer outputs a \$64\$-channel tensor at 1024x1024\$ resolution, the heavy I/O read/write times can easily stall the GPU during inference and quality assessment.

4. Zero GAN Latency at Inference

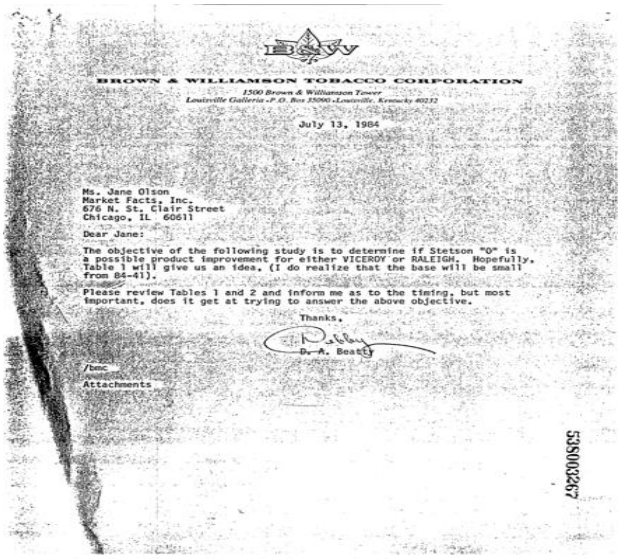
It is critical to distinguish between the Autoencoder and the GAN when reporting latency. While adding an Adversarial Loss (L_{adv}) and a Discriminator drastically increases training time and memory requirements, the GAN structure has zero impact on inference speed. During deployment, the Discriminator is completely discarded; only the Generator (the multi-level U-Net cascade) is used to process the noisy document image.

Summary of the Trade-off

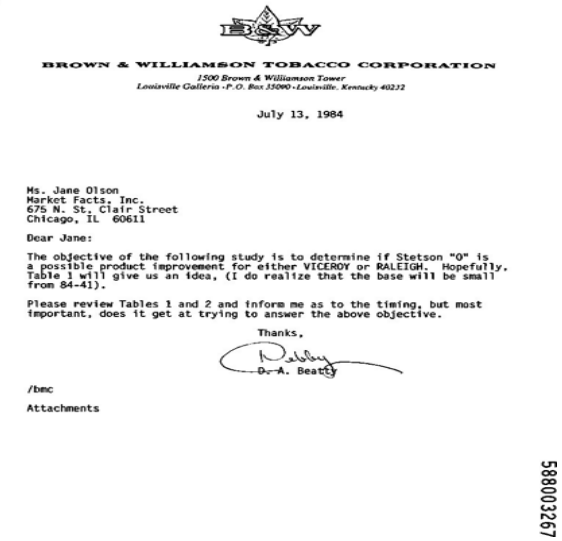
If a single U-Net takes X milliseconds to process a document, a multi-level cascade will strictly be slower, taking $X + \Delta$ milliseconds. However, this added latency (Δ) is highly optimized because the early networks process much smaller spatial dimensions. Thus, trading a slight increase in processing time for huge gains in SSIM and PSNR makes the cascaded approach highly favorable for the task.

5.7. Qualitative results

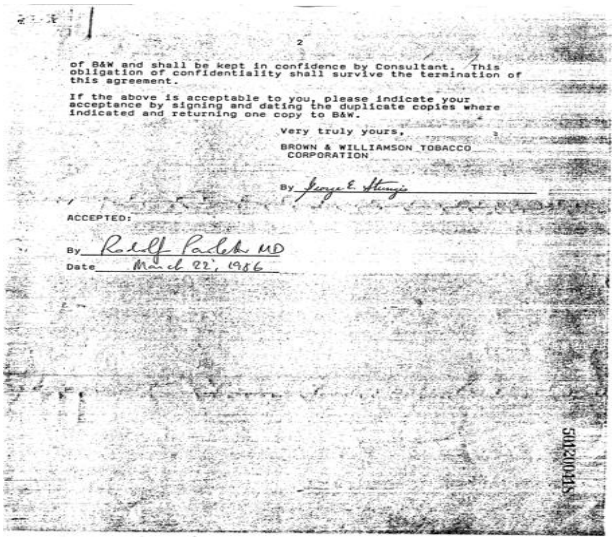
In this section we represent noisy images of different database and their corresponding cleaned output images. These different types of database contains different types of noise. Also from each database we present here different types of noise. RTobacco800 dataset contains salt and high pepper noise Restomes dibco database contains color images with degradation and FUNSD dataset contains shadow, salt and pepper noise and lines. Figure 4 shows the noisy images of different dataset and their corresponding denoised images generated by proposed architecture without distorting the content of the page. Also note that the image in (k) of figure 4 contains thick black line and handwritten components like signature and date. In the output image the black line is removed but the hand written part is preserved very well. After applying our proposed method we generate binarised output images, as the ground truth of the database is in binarised form.



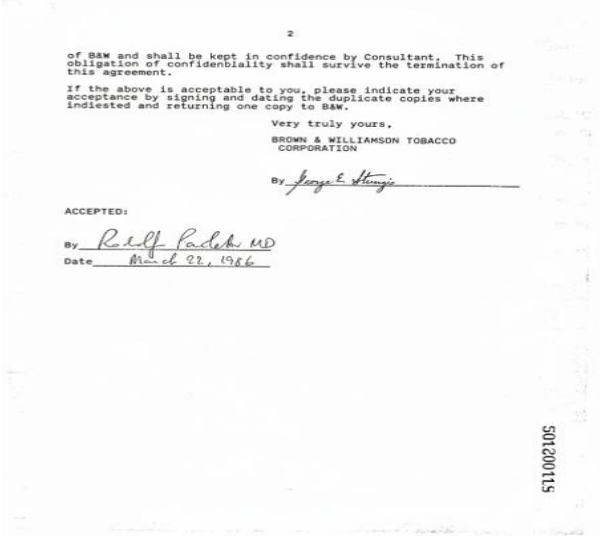
(a) R1, C1



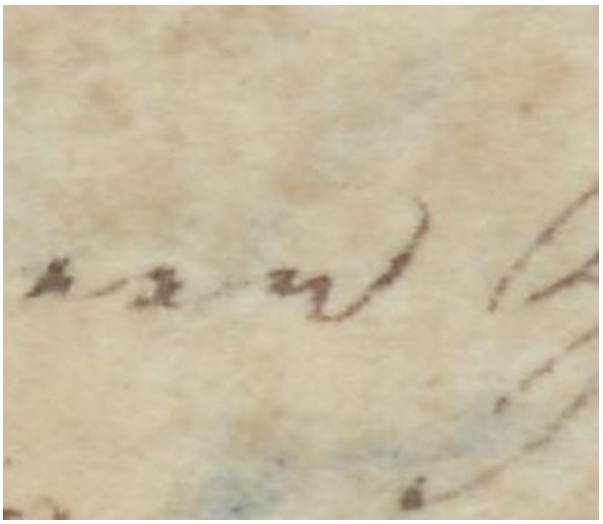
(b) R1, C2



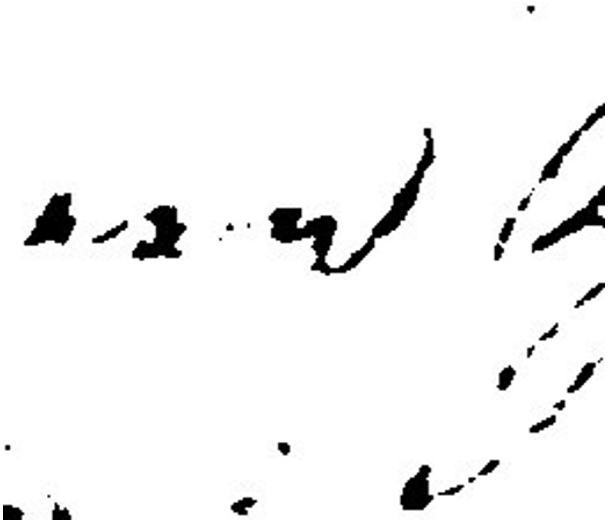
(c) R1, C3



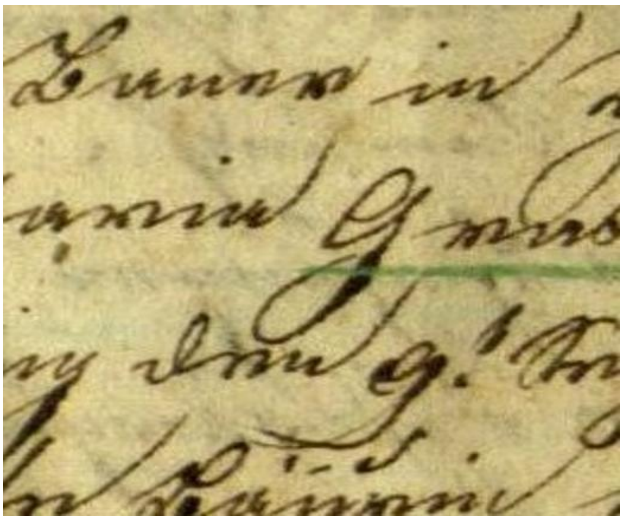
(d) R1, C4



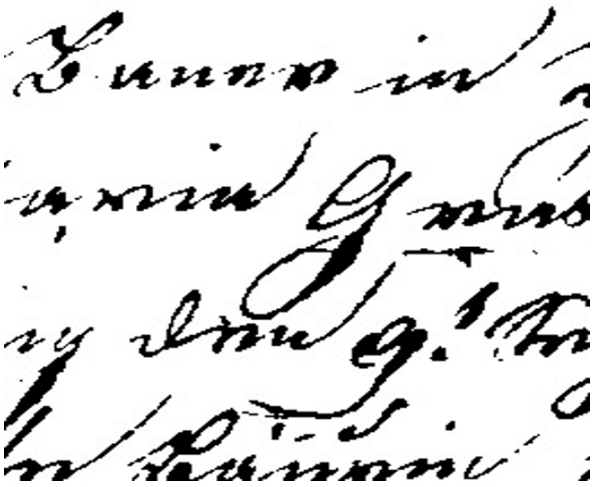
(e) R2, C1



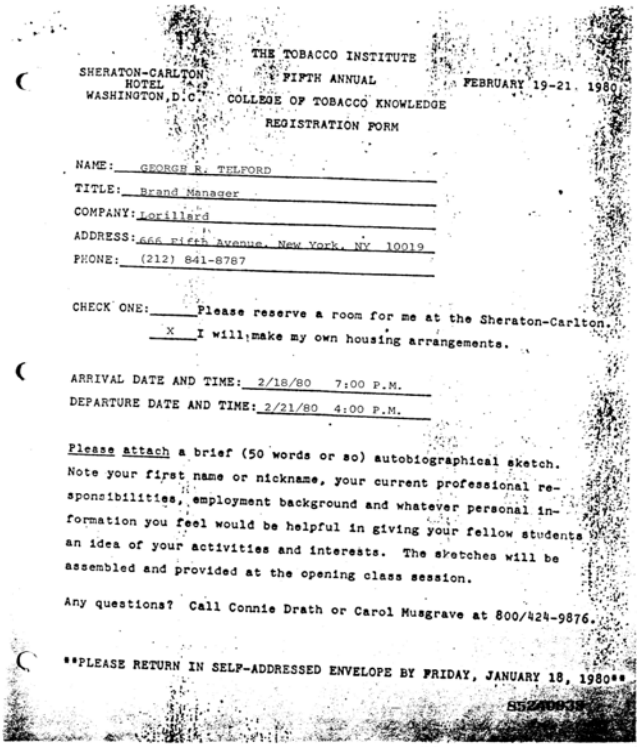
(f) R2, C2



(g) R2, C3



(h) R2, C4



THE TOBACCO INSTITUTE

SHERATON-CARLTON HOTEL WASHINGTON, D.C. FIFTH ANNUAL COLLEGE OF TOBACCO KNOWLEDGE FEBRUARY 19-21, 1980

REGISTRATION FORM

NAME: GEORGE R. TELFORD
TITLE: Brand Manager
COMPANY: Lorillard
ADDRESS: 666 Fifth Avenue, New York, NY 10019
PHONE: (212) 841-8787

CHECK ONE: ____ Please reserve a room for me at the Sheraton-Carlton.
X I will make my own housing arrangements.

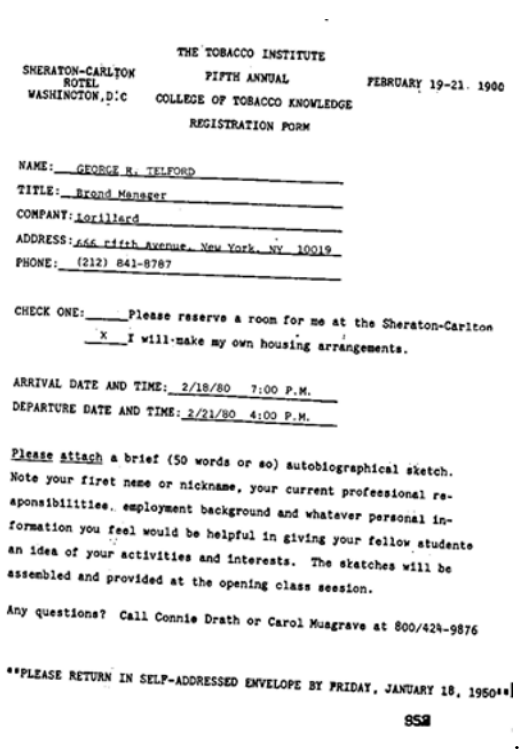
ARRIVAL DATE AND TIME: 2/18/80 7:00 P.M.
DEPARTURE DATE AND TIME: 2/21/80 4:00 P.M.

Please attach a brief (50 words or so) autobiographical sketch. Note your first name or nickname, your current professional responsibilities, employment background and whatever personal information you feel would be helpful in giving your fellow students an idea of your activities and interests. The sketches will be assembled and provided at the opening class session.

Any questions? Call Connie Drath or Carol Musgrave at 800/424-9876.

PLEASE RETURN IN SELF-ADDRESSED ENVELOPE BY FRIDAY, JANUARY 18, 1980

(i) R3, C1



THE TOBACCO INSTITUTE

SHERATON-CARLTON HOTEL WASHINGTON, D.C. FIFTH ANNUAL COLLEGE OF TOBACCO KNOWLEDGE FEBRUARY 19-21, 1980

REGISTRATION FORM

NAME: GEORGE R. TELFORD
TITLE: Brand Manager
COMPANY: Lorillard
ADDRESS: 666 Fifth Avenue, New York, NY 10019
PHONE: (212) 841-8787

CHECK ONE: ____ Please reserve a room for me at the Sheraton-Carlton
X I will make my own housing arrangements.

ARRIVAL DATE AND TIME: 2/18/80 7:00 P.M.
DEPARTURE DATE AND TIME: 2/21/80 4:00 P.M.

Please attach a brief (50 words or so) autobiographical sketch. Note your first name or nickname, your current professional responsibilities, employment background and whatever personal information you feel would be helpful in giving your fellow students an idea of your activities and interests. The sketches will be assembled and provided at the opening class session.

Any questions? Call Connie Drath or Carol Musgrave at 800/424-9876

PLEASE RETURN IN SELF-ADDRESSED ENVELOPE BY FRIDAY, JANUARY 18, 1980

(j) R3, C2

CENTER FOR INDOOR AIR RESEARCH
1099 Wintesson Road Suite 280 Linthicum, Md. 21090
(410) 684-5777 fax (410) 684-3729

APPLICATION FOR RESEARCH CONTRACT

1. PRINCIPAL INVESTIGATOR. Name, title, telephone # and mailing address.
(a) Steven R. Kleeberger, Ph.D. (b) Associate Professor (c) (410) 955-3515/955-0299
(d) Environmental Health Sciences (e) Johns Hopkins University, School of Hyg. & Pub. Hlth.
(f) 615 North Wolfe Street, Baltimore, Maryland (g) 21205

2. PROJECT TITLE. Mechanisms of Chronic Ozone Exposures: Role of Inflammation

3. KEY WORDS. Please provide three (3) key words which will be used as reference headings. Ozone, Inflammation, Mast Cell

4. INSTITUTION. Name and address of institution responsible and accountable for disposition of funds awarded on the basis of this application.
(a) Johns Hopkins University (b) 615 North Wolfe Street
(c) Baltimore (d) Maryland 21205

5. LOCATION. List location where research will be conducted if other than institution identified in #4 above.
(a)
(b)

6. INCLUSIVE DATES AND TOTAL COSTS of the specific project related to each 12 month period if more than one year is required to complete project. Summaries from budget page, item 12(i). It must be understood that awards for 2nd and 3rd periods are dependent on Center approval of continuation application.

	Inclusive Date		Total Cost
(a) 1st 12 month period	01/01/94 thru	12/31/94	$ 210,910
(b) 2nd 12 month period if required	01/01/95 thru	12/31/95	$ 212,481
(c) 3rd 12 month period if required	01/01/96 thru	12/31/96	$ 220,416

7. INSTITUTIONAL OFFICER. Name, title and telephone number of individual authorized to sign for the institution identified in #4 above. It is understood that the officer, in applying for a contract, has read and found acceptable the Center's Management of Research Contracts and Contract Administration Policy.
(a) Alan H. Goldberg, Ph.D. Assoc. Dean for Research
(c) (410) 955-9253 (d) 5/26/93

8. PRELIMINARY STUDIES*.
(a) Feasibility of proposed research
(b) Qualifications of investigator

9. EXPERIMENTAL PLAN*.
(a) Design
(b) Methods
(c) Analysis of data
(d) Interpretation of results
(e) Timetable for the investigation
(f) Literature cited

10. AVAILABLE FACILITIES AND RESOURCES.

11. OTHER SUPPORT

* Append as much material as required. Type, single space, use 8-1/2" x 11" white paper and label each sheet with name of the principal investigator in the upper right hand corner and page number at the bottom. Consecutively number each addendum beginning with page 3. Do not insert pages between pages 1 and 6, e.g., 2a, 2b, 3a, etc. Include nine copies and an original. If bonding photographs, include 2 original sets. Note: Each of the nine copies must be placed in a binder per mailing instructions.

87594142

(k) R3, C3

CENTER FOR INDOOR AIR RESEARCH
1099 Wintesson Road Suite 280 Linthicum, Md. 21090
(410) 684-5777 fax (410) 684-3729

APPLICATION FOR RESEARCH CONTRACT

1. PRINCIPAL INVESTIGATOR. Name, title, telephone # and mailing address.
(a) Steven R. Kleeberger, Ph.D. (b) Associate Professor (c) (410) 955-3515/955-0299
(d) Environmental Health Sciences (e) Johns Hopkins University, School of Hyg. & Pub. Hlth.
(f) 615 North Wolfe Street, Baltimore, Maryland (g) 21205

2. PROJECT TITLE. Mechanisms of Chronic Ozone Exposure: Role of Inflammation

3. KEY WORDS. Please provide three (3) key words which will be used as reference headings. Ozone, Inflammation, Mast Cell

4. INSTITUTION. Name and address of institution responsible and accountable for disposition of funds awarded on the basis of this application.
(a) Johns Hopkins University (b) 615 North Wolfe Street
(c) Baltimore (d) Maryland 21205

5. LOCATION. List location where research will be conducted if other than institution identified in #4 above.
(a)
(b)

6. INCLUSIVE DATES AND TOTAL COSTS of this specific project related to each 12 month period if more than one year is required to complete project. Summaries from budget page, item 12(i). It must be understood that awards for 2nd and 3rd periods are dependent on Center approval of continuation application.

	Inclusive Date		Total Cost
(a) 1st 12 month period	01/01/94 thru	12/31/94	$ 210,910
(b) 2nd 12 month period if required	01/01/95 thru	12/31/95	$ 212,481
(c) 3rd 12 month period if required	01/01/96 thru	12/31/96	$ 220,416

7. INSTITUTIONAL OFFICER. Name, title and telephone number of individual authorized to sign for the institution identified in #4 above. It is understood that the officer, in applying for a contract, has read and found acceptable the Center's Management of Research Contracts and Contract Administration Policy.
(a) Alan H. Goldberg, Ph.D. Assoc. Dean for Research
(c) (410) 955-9253 (d) 5/26/93

8. PRELIMINARY STUDIES*.
(a) Feasibility of proposed research
(b) Qualifications of investigator

9. EXPERIMENTAL PLAN*.
(a) Design
(b) Methods
(c) Analysis of data
(d) Interpretation of results
(e) Timetable for the investigation
(f) Literature cited

10. AVAILABLE FACILITIES AND RESOURCES.

11. OTHER SUPPORT

* Append as much material as required. Type, single space, use 8-1/2" x 11" white paper and label each sheet with name of the principal investigator in the upper right hand corner and page number at the bottom. Consecutively number each addendum beginning with page 3. Do not insert pages between pages 1 and 6, e.g., 2a, 2b, 3a, etc. Include nine copies and an original. If bonding photographs, include 2 original sets. Note: Each of the nine copies must be placed in a binder per mailing instructions.

87594142

(l) R3, C4

Figure 4: The noisy inputs and corresponding cleaned outputs. R1 for tobacco800 database, R2 for restomerdibco database and R3 for FUNSD database

VI. CONCLUSION

In this paper, we present a multilevel GAN and autoencoder based denoising method for multi documents of different types of noises. This proposed method don't focus on any specific document type or any specific noise type, rather it focuses on general noisy document images. For the proposed network model we calculate the loss function and have shown through different ablation studies that the loss function is best suited for proposed model. Table 2 shows results for Dataset I & II where we can see average and maximum improvement is better of the proposed model with respect to Autoencoder based and GAN based model. In the table 3 the results on Dataset 3 is shown where along with average and maximum improvement the PSNR and SSIM results are also batter with respect to Autoencoder based and GAN based model. The complexity of the proposed model is low and versatile in nature to remove noise.

REFERENCES

[1] Restomer dibco dataset.https://www.kaggle.com/datasets/lengocdatk16hcm/restomer-dibco-dataset. Accessed: 2025-11-25.

[2] Jiang, B., Li, J., & Lu, G. (2023). Enhanced frequency fusion network with dynamic hash attention for image denoising. Information Fusion, 94, 101684.

[3] Buades, A., Coll, B., & Morel, J.-M. (2005). A review of image denoising algorithms, with a new one. Multiscale Modeling & Simulation, 4(2), 490–530.

[4] Yang, C., Liang, L., & Su, Z. (2023). Real-world denoising via diffusion model.

[5] Guan, S., Lin, M., Xu, C., Liu, X., Zhao, J., Fan, J., Xu, Q., & Greene, D. (2025). PreP-OCR: A complete pipeline for document image restoration and enhanced OCR accuracy. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 15413–15425), Association for Computational Linguistics.

[6] Vaksman G. Elad M., Kawar B. (2023) Image denoising: The deep learning revolution and beyond—a survey paper. arXiv preprint arXiv:2301.03362.

[7] Gangeh, M. J., Plata, M., Motahari, H., & Duffy, N. P. (2021). End-to-end unsupervised document image blind denoising. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 7868-7877).

[8] Zheng, Q., Doermann, D., & Hancock, I. E. G. (2007). Multi-scale structural saliency for signature detection. In 2007 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-8). IEEE.

[9] Kim, K. G., & Lee, H. J. (2019). Image denoising with conditional generative adversarial networks (CGAN) in low dose chest images. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 945, 162602.

[10] Ilesanmi, A. E., & Ilesanmi, O. S. (2021). Methods for image denoising using convolutional neural network: A review. Complex & Intelligent Systems, 7(5), 2179–2198.

[11] Jaume, G., Ekenel, H. K., & Thiran, J. P. (2019). FUNSD: A dataset for form understanding in noisy scanned documents. In 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW) (Vol. 2, pp. 1-6). IEEE.

[12] Liu, K., Du, X., Liu, S., Zheng, Y., Wu, X., & Jin, C. (2023). DDT: Dual-branch deformable transformer for image denoising. In 2023 IEEE International Conference on Multimedia and Expo (ICME) (pp. 2765–2770). IEEE.

[13] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2536-2544).

[14] Xie G, Zhang Y., Wang Q.,Liu H.(2021) " Image denoising using an improved generative adversarial network with wasserstein distance." in: 40th Chinese Control Conference, CCC,, pp. 7027–7032.

[15] Brox T. Ronneberger O., Fischer P. (2015), U-net: Convolutional networks for biomedical image segmentation. IEEE Transactions on Image Processing.

[16] Anwar, S., & Barnes, N. (2019). Real image

denoising with feature attention. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3155-3164).

[17]Sadiq M., Zhu M., Chen S., Shi S D,. (2019), Image denoising via generative adversarial networks with detail loss. in: Proceedings of the 2nd International Conference on Information Science and Systems.

[18]Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., & Guo, B. (2022). Vector quantized diffusion model for text-to-image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10696–10706).

[19]Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., & Yang, M.-H. (2022). Restormer: Efficient transformer for high-resolution image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 5718-5729).

[20]Jiang, Y., Wronski, B., Mildenhall, B., Barron, J. T., Wang, Z., & Xue, T. (2022). Fast and high-quality image denoising via malleable convolutions. In Computer Vision – ECCV 2022 (pp. 420–436). Springer.

[21]Peng Y, Liu S. Wu X. Zhang Y. Wang X., Zhang L. (2019). Dilated residual networks with symmetric skip connection for image denoising. Neurocomputing, 349, 268-277.

[22]Anwar, S., & Barnes, N. (2019). Real image denoising with feature attention. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3155-3164).

[23]Xia, Z., & Chakrabarti, A. (2020). Identifying recurring patterns with deep neural networks for natural image denoising. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2814–2823.

[24]Zhang, K., Zuo, W., Chen, Y., Meng, D., & Zhang, L. (2017). Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. IEEE Transactions on Image Processing, 26(7), 3142-3155.

[25]Zhang, K., Zuo, W., & Zhang, L. (2018). FFDNet: Toward a fast and flexible solution for CNN-based image denoising. IEEE Transactions on Image Processing, 27(9), 4608-4622.