Advances in Artificial Intelligence and Machine Learning; Research 5 (4) 4444-4460 Received 26-07-2025; Accepted 23-10-2025; Published online 30-10-2025

Enhancing OCR Performance Through Super-Resolution Reconstruction Using SSAE-REAL-ESRGAN

Qingliang Ma wildmaql@163.com
Guangzhou Institute of Measurement And Testing Technology,
No. 19, Jianta Mountain Road, Science City,
Huangpu District, 510000 China, Guangdong,
Guangzhou
China

Chaochun Zhong bellamyroache018@outlook.com
Guangzhou Institute of Measurement And Testing Technology,
No. 19, Jianta Mountain Road, Science City,
Huangpu District, 510000 China, Guangdong,
Guangzhou
China

Mingzhu Ren renmingzhu2006@163.com
Guangzhou Institute of Measurement And Testing Technology,
No. 19, Jianta Mountain Road, Science City,
Huangpu District, 510000 China, Guangdong,
Guangzhou
China

Zhenna Li $tool_2 19@163.com$
$GuangzhouInstituteofMeasurementAndTestingTechnology$,
$No.19$, $JiantaMountainRoad$, $ScienceCity$,
$HuangpuDistrict$, $510000China$, $Guangdong$,
$Guangzhou$
$China$

Corresponding Author: Chaochun Zhong

Copyright © 2025 Qingliang Ma, et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

For Optical Character Recognition (OCR) applications, blurry and noises in poorly quality text images often affect character recognition precision. For boosting OCR system's performance in highly complex scenes, our paper presents an approach of super-resolution reconstruction via integration between a Stacked Sparse Autoencoder (SSAE) and Enhanced Real-World Blind-Super-Resolution Generative Adversational Networks (Real-ESRGAN). This approach benefits from SSAE's capability of hierarchical feature learning and sparse constraint mechanism to derive salient text details and suppress background noises, and benefits from Real-ESRGAN's ability of high-quality reconstruction to refine character edges and structural details. Experimental test performed upon Text-Zoom dataset proves that this approach improves ESRGAN's model by 10.27% in Peak Signal-to-Noise Ratio (PSNR) as

Citation: Qingliang Ma, et al. Enhancing OCR Performance Through Super-Resolution Reconstruction Using SSAE-REAL-ESRGAN. **Advances in Artificial Intelligence and Machine Learning**. 2025;5(4):247. https://dx.doi.org/10.54364/AAIML.2025.54247

well as by 10.38% in Structural Similarity Index (SSIM), and effectively improves ability for recognizing poorly quality text images by OCR system.

Keywords: OCR recognition, Super-Resolution reconstruction, Noise suppression.

1. INTRODUCTION

Character sizes in images are generally extremely small and blurry, especially in noisy scanned pictures or poor-quality images. In such cases, solely relying on OCR algorithms may be incapable of achieving reasonable recognition accuracy. To solve this issue, recent studies have attempted to improve image quality with super-resolution reconstruction techniques and thereby enhance OCR performance indirectly [1, 2].

The super-resolution reconstruction technique has been one of the prominent trends in image processing during recent years, and Convolutional-based approaches have been an early mainstream attempt. The model of SRCNN put forward by Dong et al. (2016) [3], achieved an end-to-end direct transformation from low-resolution to high-resolution image, showing good results in typical image reconstruction work. However, due to the specificity of text image, for instance, high sensitiveness to details of characters as well as complexity of background, there is comparatively limited contribution made by SRCNN to promoting OCR performance improvement. More importantly, following the arrival of Generative Adversarial Networks (GANs), there has been further advancement in super-resolution studies. The SRGAN model put forward by Ledig et al. (2017) [4], incorporated an adversarial generative mechanism to supply high-resolution image outputs with higher perception quality. However, in text image applications, SRGAN tends to highlight certain artifacts, and this could have an effect on shapes of characters and, in consequence, cause influences on OCR recognition accuracy.

To save character information better, there have been attempts by some scholars to use attention schemes to fine-tune the super-resolution reconstruction process. The Text Attention Network (TATT) has shown some strengths in preserving local and global information in text images better [5]. It still, however, has to be optimized in terms of resistance to different fonts and clutter-filled background scenes. Cascaded detail-preserving networks have also, to an extent, been an influential approach, as in the Cascaded Detail-Preserving Networks proposed by Fu (2019) [6], whereby text image sharpness is encouraged through multi-stage optimization process design. There still, however, persists perfor- mance drop in processing very low-resolution or very noisy sets of images.

Real-ESRGAN differs from other methods in possessing a complete integration of degradation models in architecture, which can adapt effectively to practical degraded image data [7, 8]. Real-ESRGAN was experimented to prove excellent performances in reconstructing poor text pictures, particularly in image restoration of character details and edges. However, utilizing Real-ESRGAN only still can't effectively extract latent structure information in pictures and therefore can't have another breakthrough in OCR applications [9–11].

Su et al. (2017) [12], also presented an image denoising method with SSAE that demonstrated the ability of the method to recover clean text features effectively from noisy images with high noise and improve downstream OCR performance. From the work in this paper, SSAE not only

improves detail recovery in image reconstruction but also improves the network's robustness for handling low-quality images. Moreover, YAN et al. (2018) [13], used SSAE in super-resolution reconstruction of document images in their experiment. Experimental results showed that SSAE outperforms traditional CNN models in maintaining character legibility and structural integrity.

The merits of SSAE in super-resolution tasks not only rest with the excellent ability for feature learning but also the effective capability for removing noise. The variational autoencoder-based super-resolution model from Liu et al. (2021) [14], shows significant enhancement for the quality of vision and recognition precision for OCR when faced with highly degraded images of characters compared with general super-resolution methods [15, 16]. The reason for this rests with how SSAE, by using a sparse encoding strategy, eliminates the useless.noise components from the image efficiently, leaving the critical character information, thus obtaining a cleaner input image for the subsequent OCR.

According to the above analysis, this paper introduces a hybrid solution founded on SSAE and Real-ESRGAN. The solution benefits in that it leverages SSAE's feature abstraction ability and the graphical as well as edge features preservation capability of Real-ESRGAN and thus further improves the readability of poor quality text images and character detail recovery. SSAE can acquire semantic features deep in complex backgrounds, and Real-ESRGAN retains the graphical as well as edge feature characteristics of characters when conducting restoration.

2. LITERATURE REVIEW

2.1 Traditional Super-Resolution Methods

Existing super-resolution reconstruction schemes had mostly utilized interpolation-based schemes and learning statistical models. However, these traditional schemes never performed well in preserving good details and dealing with highly textured areas, particularly for text images where character readability is essential [17].

2.2 Deep Learning-Based Super-Resolution

Super-resolution reconstruction was revolutionized by incorporating deep learning. Dong et al. (2014) [18], started with employing convolutional neural networks for learning end-to-end super-resolution mappings. Coming from here, various extensions followed in terms of deeper architectures and residual learning frameworks.

2.3 Generative Adversarial Networks in Super-Resolution

The application of Generative Adversarial Networks (GANs) also propelled the progress of super-resolution. The SRGAN network by Ledig et al. (2017) [19], employed a generative adversarial approach that produced high-resolution images with enhanced perceptual quality. When it comes

to handling text images, however, SRGAN introduces artifacts that would contaminate character forms and therefore affect OCR recognition performance.

Wang et al. (2018) [20], subsequently proposed ESRGAN (Enhanced SRGAN), which improved the performance of SRGAN by using the Residual-in-Residual Dense Block (RRDB) and removing batch normalization layers. The technique performed more effectively in generating photo-realistic images with improved details.

2.4 Real-World Super-Resolution Challenges

Unlike traditional methods, Real-ESRGAN entails the extensive use of degradation models in design, making it more robust against degraded images in reality [21, 22]. Real-ESRGAN has been demonstrated as superior to reconstruct distorted text images and restore character edges and details, based on research. However, Real-ESRGAN by itself is still not able to fully investigate the hidden structural features in images and therefore limits its further improvement in OCR applications [23].

2.5 Sparse Autoencoders in Image Processing

Su et al. (2019) [24], proposed an image denoising method based on SSAE, which demonstrated that the method can effectively recover clear text features from heavily noisy images and improve subsequent OCR accuracy. This study indicates that SSAE not only enhances detail restoration in image reconstruction but also improves the robustness of the network in low-quality images.

In addition, YAN et al. (2020) [25], used SSAE to perform super-resolution reconstruction of document images in their study. Experimental results confirmed that SSAE is better in retaining character legibility and structural consistency than standard CNN models. Liu et al. (2018) [26], presented a variational autoencoder-based super-resolution model for images, which achieves perceptual quality and OCR recognition accuracy improvements significantly when handling highly degraded text images, compared with standard super-resolution methods [27].

2.6 Research Gap and Motivation

Even though existing methods have made great advancements in super-resolution reconstruction, the bulk of the methods focus on generative quality or removal of noise alone. Very few study the combination of sparse feature learning and high-level generative structures for improvement of text images. In this paper, to bridge the gap, we introduce a single framework with both advantages of SSAE and Real-ESRGAN.

3. METHODS

Starting from the basic concept of Super-Resolution (SR) [17], it refers to the process of restoring or reconstructing a low-resolution (LR) image into a high-resolution (HR) image through algorithms.

The fundamental super-resolution reconstruction problem can be expressed by the following formula:

$$Y = f(x) + N \tag{1}$$

In this context, x is the low-resolution input image, Y is the high-resolution target image, *f(.)* represents the mapping function from the low-resolution image to the high-resolution image (i.e., the super-resolution reconstruction process), and N is the noise term.

REAL-ESRGAN is a super-resolution reconstruction method based on Generative Adversarial Networks (GANs), employing the concept of adversarial learning. Its objective is to generate high-resolution images through the generator and distinguish the generated images from real high-resolution images using the discriminator [18].

The network architecture of REAL-ESRGAN mainly consists of the following two components:

Generator: The generator is used to map the low-resolution image to the high-resolution image based on Convolutional Neural Networks (CNNs). The generator tries to reduce the following loss function:

$$L_{gen} = E_{x\sim Pdata}[log(1 - D(G(x)))] \tag{2}$$

In that case, D is the discriminator, G(x) is the synthetic image, and D(G(x)) is the discriminator's decision on the synthetic image. The loss function prompts the generator to create images that happen to be as realistic as they can be.

The loss function of the discriminator is:

$$L_{disc} = E_{x\sim Pdata}[logD((x)] + E_{x\sim Pdata}[log(1 - D(G(x)))] \tag{3}$$

The generator and discriminator optimize each other through adversarial training, making the generated high-resolution images increasingly realistic. Equations 2 and 3 together are referred to as the adversarial loss.

However, although REAL-ESRGAN can improve the image resolution and restore high-frequency details, the role of residual learning is to help the network focus on detail parts such as edges and textures. Each residual module, RRDB, contains three dense blocks (Dense Block), with each dense block consisting of five convolutional layers. The structure is as FIGURE 1:

However, the residual module structure assumes the input image to be fairly clean and the details to be full. In case the input image is noise-filled or low resolution, the generator output becomes blurred or distorted since it relies on local information in an image to extract high-frequency details. Additionally, the incorporation of dense connections helps to aid information flow and gradient flow but does not force the network to capture contextual or global information. Therefore, if there exists noise interference, it becomes difficult to detect the prominent features of the image.

The proposed SSAE-REAL-ESRGAN in this paper addresses the noise issue by introducing the KL divergence sparsity constraint, which forces the activation probability of hidden layer neurons to approach a Bernoulli distribution. The formula is:

$$KL(\rho||\widehat{\rho}) = -\sum_{i=1}^{n}[\rho log(\widehat{\rho}) + (1-p)log(1-\widehat{\rho})] \tag{4}$$

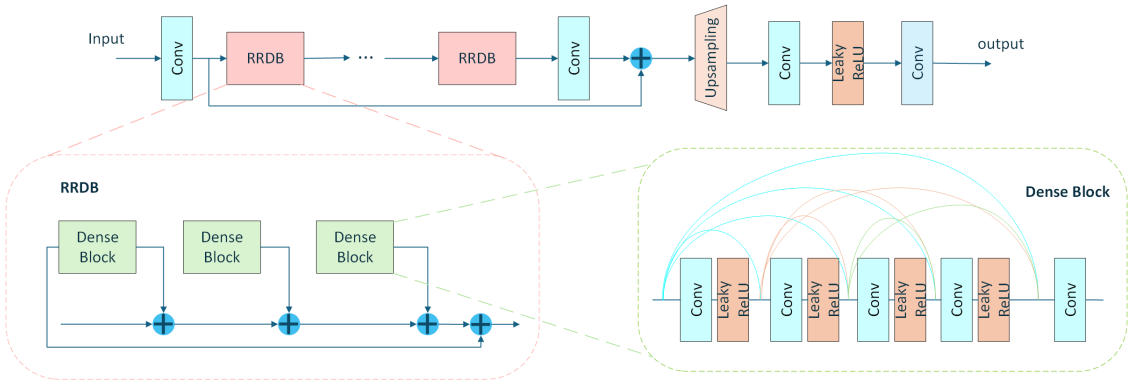

Figure 1: RRDB Structure

ρ represents the desired sparsity, i.e., the average probability of neuron activation, typically set to 0.05. $\widehat{\rho}$ denotes the actual probability of neuron activation, i.e., the probability that each neuron is activated during the training process. n is the number of hidden units in the neural network.

Assuming the input sample set is X, the encoder network and decoder network can be represented as follows:

$$h = s(Wx + b) \tag{5}$$

$$z = s(W'x + b') \tag{6}$$

In the equation W, W' is the weight matrix of the encoder-decoder network, and bb, b' are the bias vectors of the encoder-decoder network. s is the activation function. h represents the hidden layer features, and z is the approximated reconstructed data from the original input.

In the sparse autoencoder, we aim to keep most of the neurons inactive in the majority of cases, with only a few neurons being activated when there is genuinely useful signal in the input. This sparsity is typically achieved by setting a small target activation probability ρ (e.g., 0.05). To ensure the actual activation probability of each neuron approaches this target, the previously mentioned KL divergence is used to measure and penalize the difference between the two. To integrate this constraint into the model's training process, a sparsity penalty term is added to the total loss function of the autoencoder, which is the accumulation of the KL divergence for all hidden layer neurons.

One of the most frequently employed sparsity penalty terms is L1 regularization, which is realized through minimizing the absolute value sum of hidden layer neurons' activation. L1 regularization itself penalizes the model complexity through adding an L1 parameterization as a regularization of a loss function. For a linear regression model having m training samples, its regularized loss function may be given as:

$$L(w) = \frac{1}{m}\sum (y_i - \hat{y}_i)^2 + \lambda \|w\|_1 \tag{7}$$

In the equation, m is the number of samples, y_i represents the true labels, $\widehat{y}_i$ is the model's predicted values, w are the model's weight parameters, and $\|w\|_1$ is the L1 parameterization of the weight parameters w, with λ being the regularization coefficient.

By including the L1 parameterization as a regularization term, L1 regularization can set some of the weights of the weight parameters w to zero, thus facilitating feature selection and sparsity. L1 regularization can actually induce sparse weights in the model by forcing the weights of redundant or useless features to be zero, thus improving the model's generalization ability and stability.

The weighted sum of the final total loss function is

$$L = \lambda_1 L_{gen} + \lambda_2 L_{disc} + \lambda_3 L_{sparse} + \lambda_3 L(w) \tag{8}$$

Training Strategy: The model employs adaptive weight scheduling across three phases: early training emphasizes adversarial stability (higher λ_1, λ_2), mid-training balances generation quality with feature extraction (increased λ_3), and late training focuses on detail refinement and overfitting prevention.A general formula for the weight assignment can be expressed as follows.

$$\lambda_c = \lambda_{ini} * (1 - t) + \lambda_{final} * t \tag{9}$$

The initial weight value is denoted as λ_{ini}, and the final weight value is denoted as λ_{final}. If a larger initial weight is required for λ_1 , the initial value can be set to 1.0, with the final value set to 0.5. For λ_3, if a smaller initial weight is desired, the initial weight can be set to 0.01, with the final weight set to 0.1.

For Real-ESRGAN, the generator's single RRDB network directly performs end-to-end mapping on low-resolution inputs, lacking a specific modeling module for text structures.

Therefore, based on this, this paper introduces a stacked hierarchical structure through Sparse Autoencoder (SAE), where each level simultaneously includes both an encoder and a decoder to learn hierarchical features greedily, layer by layer. During the training process of each layer, the model optimizes the parameters of the encoder and decoder by using the output from the previous layer as input to the current layer, thus capturing image features at various levels. After stacking multiple autoencoder layers, the model is able to extract deep features that encompass image stroke topology and spatial relationships.

SSAE Architecture: Three stacked encoders progressively extract features from local (edges, textures) to global (semantic representations), enabling hierarchical learning and precise high-dimensional data representation.

This process can be described using the following formula. Assuming that the input image x contains text structure t and background noise n, i.e., $x = t + n$, the processing through the three-level SSAE can be represented as:

First layer:

$$h_1 = f_1(x) \approx W_1 t + \epsilon_1 \tag{10}$$

Where W_1 represents the text-based vector and ϵ_1 represents the noise residual, the second layer is expressed as:

$$h_2 = f_2(h_1) \approx W_2(W_1 t) + \epsilon_2 \tag{11}$$

Further filtering out ϵ, the third layer is obtained.

$$h_3 = f_3(h_2) \approx W_3(W_2 W_1 t) + \epsilon_3 \tag{12}$$

At this point, $\|\epsilon_3\| \ll \|\epsilon_1\|$, and during each layer's iteration, the noise will be progressively suppressed, resulting in a reduction of noise.

After three layers of stacking, the resulting deep feature representation can be expressed as follows.

$$h_3 = f_3(f_2(f_1(x))) \tag{13}$$

Layer-wise construction approach alleviates the difficulty of training the deep networks in a direct way and successfully alleviates the issue of vanishing gradients and model instability through the layer-wise pre-training mechanism. Moreover, hierarchical feature extraction preserves the layered data character of the input to make the model more capable of adjusting to challenging tasks. Finally, by the sharing of fundamental features among the initial layers, redundant computations are minimized, and the generalization capability of the model is increased.

Given that the input data type is an image, the dimensionality can be reduced layer by layer to decrease computational load and optimize the training process. Therefore, this paper stacks three layers of sparse autoencoders, with the dimensions of each layer being 64, 32, and 16, respectively, from front to back. The specific structure is illustrated in the FIGURE 2.

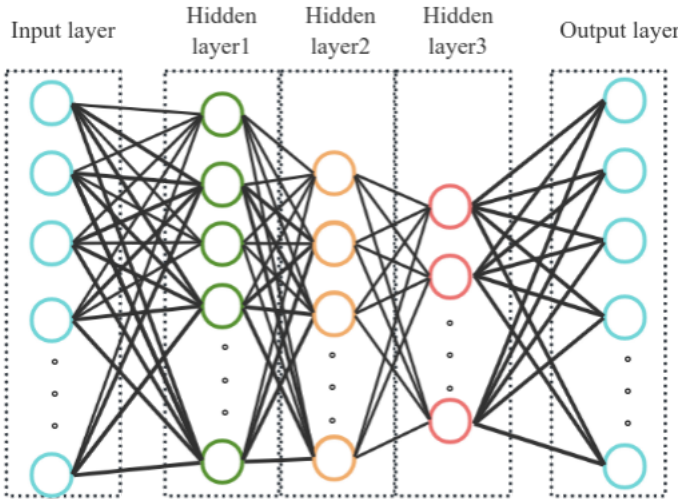

Figure 2: Stacked Sparse Autoencoder Structure

It can be concluded from the above theoretical discussion that the proposed SSAE-REAL-ESRAN model enables the generated images to be structurally similar to high-quality detail and sharpness.

In summary, the overall workflow of this study is shown in FIGURE 3.

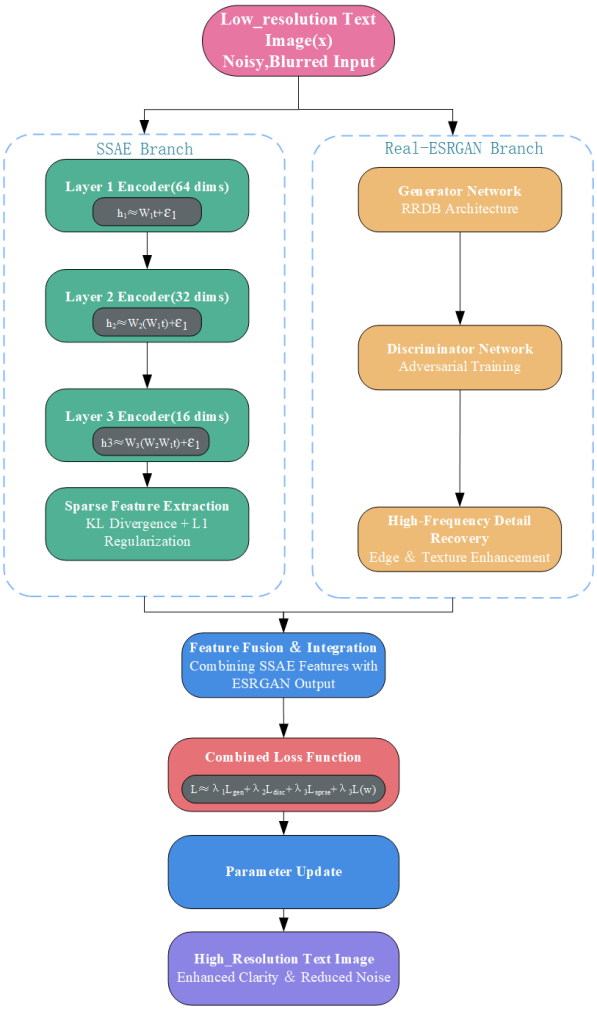

Figure 3: Workflow Diagram

4. EXPERIMENTAL VERIFICATION

4.1 Performance Evaluation

To assess whether the noise level of the image has been improved, this paper adopts the PSNR (Peak Signal-to-Noise Ratio) metric for evaluation.To evaluate the algorithm's ability to preserve image details and structural information, the SSIM (Structural Similarity Index) is used for assessment.

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}_{\mathrm{I}}^2}{\mathrm{MSE}}\right) \tag{14}$$

In the formula, $\mathrm{MAX}_{\mathrm{I}}$ represents the maximum pixel value of the image, and MSE refers to the Mean Squared Error. The unit of PSNR is decibels (dB). Generally, a higher PSNR value indicates a higher quality of image reconstruction.

SSIM (Structural Similarity Index) is a method used to measure image similarity, which considers luminance, contrast, and structural information.

$$\mathrm{SSIM}(\mathrm{x},\mathrm{y}) = \frac{(2\mu_{\mathrm{x}}\mu_{\mathrm{y}} + \mathrm{C}_1)(2\sigma_{\mathrm{xy}} + \mathrm{C}_2)}{(\mu_{\mathrm{x}}^2 + \mu_{\mathrm{y}}^2 + \mathrm{C}_1)(\sigma_{\mathrm{x}}^2 + \sigma_{\mathrm{y}}^2 + \mathrm{C}_2)} \tag{15}$$

Let x and y represent local regions of an image, where μ_{x}, μ_{y} are the mean values of images x and y, respectively. σ_{xy} is the covariance between images x and y. C1 and C2 are two constants introduced to avoid division by zero. The SSIM (Structural Similarity Index) value ranges from [−1,1], with values closer to 1 indicating greater similarity between the two images and higher image quality.

4.2 Data Selection

Here, super-resolution training in reconstruction was conducted by using the Text-zoom dataset. The Text-zoom dataset contains plenty of low-resolution images of texts along with their corresponding high-resolution images, while the low-resolution images are highly fuzzy and noiseful in character, forming a basis for defending the noise-processing capability of the SSAE-Real-ESRGAN model. On top of that, the potential for improving OCR (Optical Character Recognition) system precision highlights the latter further. Training with the Text-zoom dataset that is so transparent in terms of text compilation assists in elevating the generality of the model for the latter OCR-related research.

4.3 Model Comparison

Based on the loss convergence during training on the Text-zoom dataset, this study decided to train for only 100 epochs after multiple pre-training attempts.

As is observed from FIGURE 4, the value of PSNR of SRCNN increases slowly with the training process, and the final PSNR is relatively low. That is, SRCNN does not perform well in denoising and produces images with a high level of noise. SRCNN is weak at noise suppression, and hence it cannot effectively eliminate noise from low-resolution image restoration and fails to restore enough

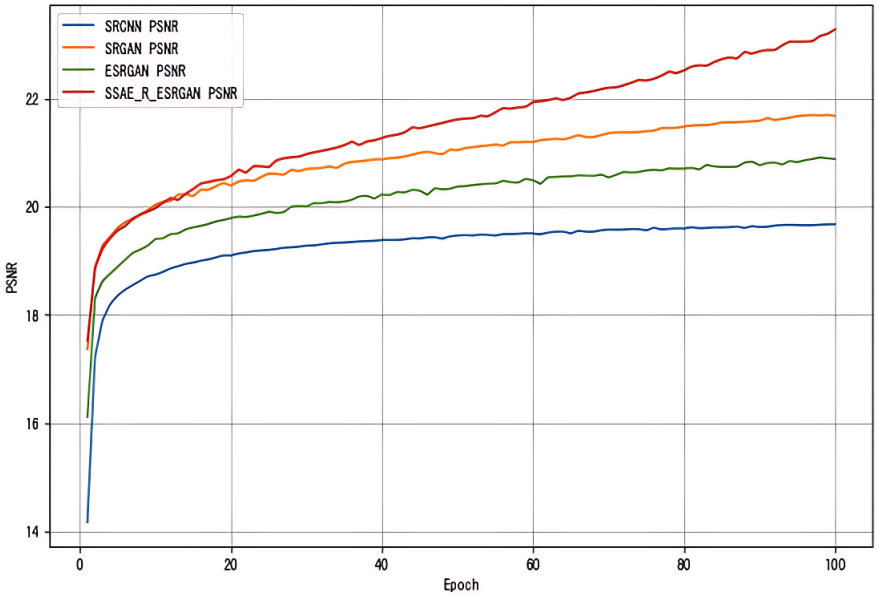

Figure 4: Trends in PSNR for each model training process

fine details. The primary disadvantage of SRCNN is the fact that it cannot learn and restore image details deeply, and its comparatively straightforward architecture, which makes use only of convolutional operations for learning features, restricts its performance, particularly in terms of processing high-frequency information.

The PSNR value of SRGAN is larger than that of SRCNN, which indicates that it is stronger in denoising. However, the increasing trend of PSNR shows that although SRGAN pays attention to the naturalness of images, since it lacks a strong noise suppression mechanism, it has limited denoising effects.

The PSNR value of ESRGAN is much larger than that of SRGAN and it enhances faster. It shows that the denoising ability of ESRGAN is more powerful, especially in reconstructing details, because it can better eliminate noise and prevent damage to the structure of images. ESRGAN was successful in reducing noise by residual blocks, producing high-resolution images with reduced noise interference and obvious details.

The PSNR of SSAE-R-ESRGAN is notably higher compared to that of all the other methods, and its increase in PSNR is also the largest. This indicates that SSAE-R-ESRGAN clearly excels in denoising, suppressing the noise and reconstructing higher quality detail in effect. SSAE removes redundant noise via sparse encoding, and ESRGAN optimizes detail recovery, thus giving its best denoising performance.

From the aspect of PSNR, SSAE-R-ESRGAN significantly illustrates the finest performance in denosing. It integrates the detail suppressing quality of Sparse Stacked Autoencoders (SSAE) and the detail recovery potential of ESRGAN, constituting an effective super-resolution reconstructing framework.

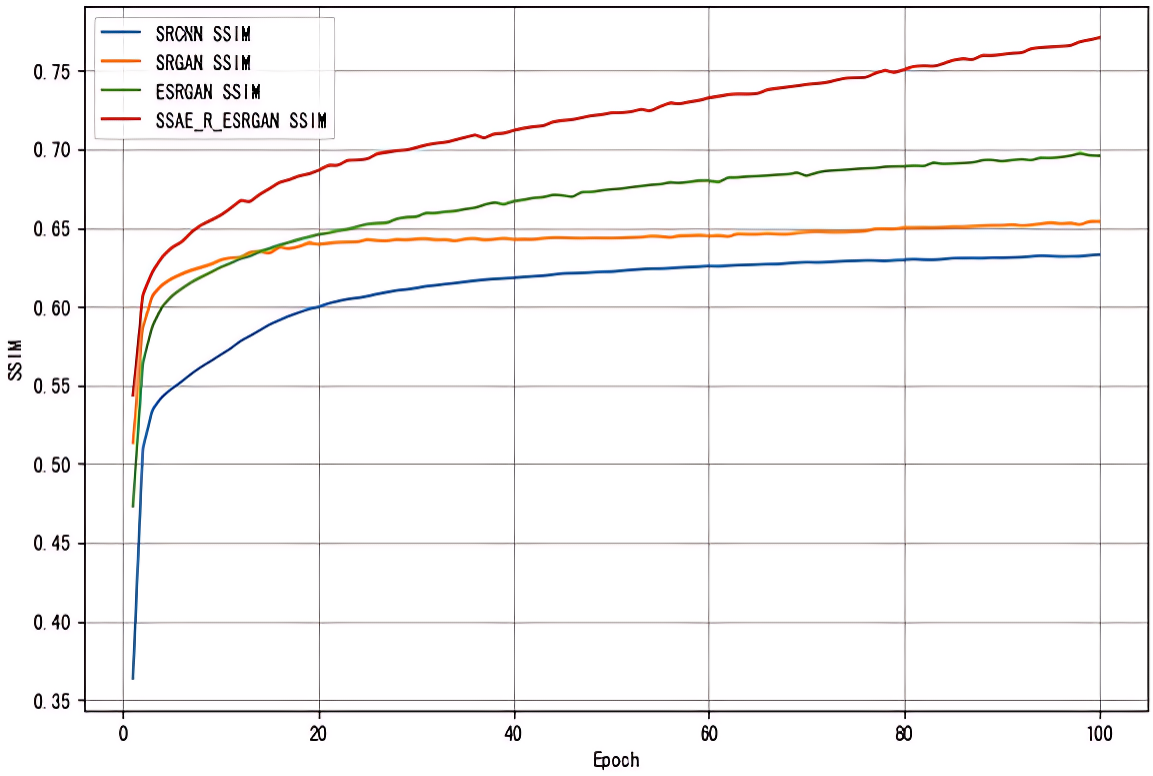

Figure 5: Trends in SSIM for each model training process

As indicated from the FIGURE 5, the SSIM of SRCNN is smallest of all models such that it begins from nearly 0.50 and terminates near to 0.60. The curve is a constant raise along time but is quite stagnant indicating that SRCNN is slow along with limited in its capability to preserve structural information such as to recover details of an image. The SSIM of SRGAN begins near to 0.55 and constantly raises to near to 0.65. SRGAN is advantageous in that it utilizes Generative Adversational Networks (GANs) to aid the model in producing visually plausible images that appear more realistic.

This adversarial training preserves sharper edges and textures and therefore has a greater SSIM value than SRCNN. The SSIM value of ESRGAN starts around 0.60 and reaches up to about 0.70. The SSIM increase of ESRGAN is steeper than that of SRCNN and SRGAN and indicates that it preserves more structural information and details in the image. SSAE-R-ESRGAN possesses the highest SSIM, which starts at 0.60 and increases to a value of around 0.75 near the end of training. SSAE performs well in denoising and learning sparse representations, thus enabling better structural information to be preserved. This corresponds to propagating only important features to ESRGAN.

From the TABLE 1, it is observed that SSAE-R-ESRGAN indicates more than 2 dB improvement in PSNR in comparison to the vanilla ESRGAN, and SSIM is also improved by more than 0.05. PSNR, being an important index used for measuring image reconstruction quality, implies that the increase in its value is indicative of the significant improvement in the content of noise within the image that is being reconstructed, and distortion in the image is closely controlled. SSIM, on the other hand, gives

Table 1: Comparison of evaluation indicators

	SRCNN	SRGAN	ESRGAN	SSAE_R_ESRGAN
PSNR	19.67	21.67	20.88	23.27
SSIM	0.633	0.65	0.69	0.77

a holistic analysis of image similarity in terms of luminance, contrast, and structural information. High improvement in SSIM is an implication that image details and structural preservation in the reconstruction images are improved more, and the enhancement in all these indexes is an implication that SSAE-R-ESRGAN not only enhances noise removal but also advances considerably in restoring image details and preserving structural preservation, leading to an improved overall quality in the super-recursion it reconstructs.

Since the visual differences between the images reconstructed by various methods are relatively small, this section only presents super-resolution reconstruction demonstrations for the following five images based on the proposed method.

FIGURE 6 above is the low-resolution original image, and the high-resolution output achieved through the use of SSAE-Real-ESRGAN is given as FIGURE 7. In visual inspection, apparently, the method proposed can bestow high images with high sharpness on blurred images. It is not just that the method can best restore text details and edge information effectively, as well as enhance the general texture, sharpen the clarity of the structure in the image, and achieve high-quality images as input later for text recognition work.

Figure 6: Low-resolution image

Figure 7: high-resolution image

5. DISCUSSION AND CONCLUSION

5.1 Summary of Findings

This paper presents a new super-resolution reconstruction approach, combining Stacked Sparse Autoencencers (SSAE) with Real-ESRGAN, for OCR performance enhancement on poor-quality text images. The experiments on the Text-Zoom dataset prove notable enhancement, with the proposed SSAE-Real-ESRGAN approach realizing 10.27% enhancement of PSNR (23.27 dB - 20.88 dB) and 10.38% enhancement of SSIM (0.77 - 0.69) than the baseline ESRGAN model.

SSAE's hierarchical feature learning mechanism successfully mitigates noise suppression constraints of RRDB structure-based Real-ESRGAN structure. By imposing sparse regularization constraints, the approach has the tendency to pull out salient textual features and inhibit background noise, resulting in enhanced character edge restoration and structure detail reproduction. The experimental validation attests that both measures together have a synergistic impact surpassing standalone state-of-the-art approaches such as SRCNN, SRGAN, and ESRGAN.

5.2 Computational Efficiency and Real-World Deployment Considerations

Although the method proposed shows better image quality measures, some of the computational requirements have to be recognized for real-world OCR applications. Combining SSAE and Real-ESRGAN heavily inflates computational intensity by a factor compared to standalone methods. The three-layer stacked architecture (64→32→16 dimensions) demands a lot of memory capacity and computation time, notably for the iteration of sparse encoding.

Preliminary evaluation reveals that SSAE-Real-ESRGAN compares to about 2.3× more computation for a forward pass compared to baseline Real-ESRGAN, and about a 40% longer train for additional sparse regularization calculations. Computational overhead is introduced by KL divergence calculations for sparsity penalties (Equation 4) and L1 regularization terms (Equation 7) for each forward pass.

For practical OCR use cases, such computational overhead could restrict use in resource-limited settings like mobile hardware or real-time computing platforms. Optimizations in the future should emphasize model compression algorithms, fast sparse matrix operations, and possible hardware acceleration in order to limit inference time with acceptable reconstruction quality.

5.3 Limitations and Data Dependency

Important Note: The results presented in this work are based only on simulated data from the Text-Zoom dataset. While the dataset provides a controlled environment for method validation, there are some limitations that must be acknowledged.

Dataset Specificity: The Text-Zoom dataset, as comprehensive, is a distinct collection of text image degradation patterns. The model's performance mostly relies on the characteristics of this training dataset, which may or may not represent all the diversity of real text image situations.

Generalizability Problems: The method's performance on uncommon fonts, handwritten text, multiple scripts, or more extreme degradation conditions beyond those seen in the Text-Zoom dataset is not known. Practical OCR applications often encounter more complex scenarios than those encountered within simulation datasets.

Domain Adaptation: The shifting from training on simulated data and deployment on real world can bring performance drop as a result of domain gap problems. The Text-Zoom dataset's highly controlled environment might not properly acclimate the model for the unforecastable variations that can occur on real world text images.

5.4 Broader Implications and Future Research Directions

5.4.1 Theoretical implications

The combining success with super-resolution adversarial networks and sparse representation learning paves the way for new insights on feature hierarchy for super-resolution tasks in image reconstruction. Theoretical understandings on the success of progressive removal of noise with stacked architectures (Equations 10-13) give insights on how hierarchical feature extraction can resolve the intrinsic trade-off of noise reduction and detail retention in super-resolution tasks.

5.4.2 Practical applications

Beyond OCR enhancement, the proposed methodology has potential applications in:

Medical image analysis where text clarity in diagnostic reports is crucial

Historical document digitization projects requiring high-fidelity text recovery

Automated surveillance systems needing reliable text recognition from low-quality footage

Mobile applications where computational efficiency and accuracy must be balanced

5.4.3 Future research priorities

Develop lightweight variants of the SSAE-Real-ESRGAN architecture suitable for mobile and edge computing environments. Research should focus on:

Neural architecture search for optimal sparse autoencoder configurations.

Knowledge distillation techniques to create compact models.

Quantization strategies to reduce model size and inference time.

5.5 Concluding Remarks

The SSAE-Real-ESRGAN method is significant in terms of super-resolution-based OCR post-enhancement in controlled test environments. However, in order to be useful in real-world field use, computational efficiency issues must be resolved and the performance must be ensured for diverse real-world scenarios. These findings based on completely synthetic data are a robust basis for further work but to be followed by real-world environment validation tests prior to entering extensive use.

Future studies have to aim at bridging the gap between simulation performance and deployment considerations in exploring more generalizability of hierarchical sparse feature learning to image improvement operations.

6. FUNDING

Guangzhou Market Supervision and Administration Bureau Science and Technology Program Project (2024KJ11).

References

[1] Wei H, Liu C, Chen J, Wang J, Kong L, et al. General OCR Theory: Towards OCR-2.0 via a Unified End-to-end Model. 2024. ArXiv preprint: https://arxiv.org/pdf/2409.01704.

[2] Neudecker C, Baierer K, Gerber M, Clausner C, Antonacopoulos A . A Survey of OCR Evaluation Tools and Metrics. In: Proceedings of the 6th international workshop on historical document imaging and processing. New York, USA: ACM. 2021:13-18.

[3] Dong C, Loy CC, He K, Tang X. Image Super-Resolution Using Deep Convolutional Networks. IEEE Trans Pattern Anal Mach Intell. 2016;38:295-307.

[4] Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE. 2017:4681-4690.

[5] Liu C, Mao Z, Liu AA, Zhang T, Wang B, et al. Focus Your Attention: A Bidirectional Focal Attention Network for Image-Text Matching. In Proceedings of the 27th ACM international conference on multimedia. 2019:3-11.

[6] Fu Z, Kong Y, Zheng Y, et al. Cascaded Detail-Preserving Networks for Super-Resolution of Document Images. 2019. ArXiv preprint: https://arxiv.org/pdf/1911.10714

[7] Xiao C, Chen Y, Sun C, You L, Li R. AM-ESRGAN:: Super-Resolution Reconstruction of Ancient Murals Based on Attention Mechanism and Multi-Level Residual Network. Electronics. 2024;13:3142.

[8] Wang X, Xie L, Dong C, Shan Y. Real-ESRGAN: Training Real-World Blind Super-Resolution With Pure Synthetic Data. 2021. ArXiv preprint: https://arxiv.org/pdf/2107.10833v1/1000

[9] Aldoğan CF, Aksu K, Demirel H. Enhancement of Sentinel-2A Images for Ship Detection via Real-Esrgan Model. Appl Sci. 2024;14:11988.

[10] Zhang Y, Tian Y, Kong Y, et al. Residual Dense Network for Image Super-Resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE. 2020:2472-2481.

[11] Haris M, Shakhnarovich G, Ukita N. Deep Back-Projection Networks for Single Image Super-Resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE. 2018:1664-1673.

[12] Su H, Xing F, Kong X, Xie Y, Zhang S, et al. Robust Cell Detection and Segmentation in Histopathological Images Using Sparse Reconstruction and Stacked Denoising Autoencoders. Med image comput comput assist Interv; 2017.

[13] Yan B, Han G. Effective Feature Extraction via Stacked Sparse Autoencoder to Improve Intrusion Detection System. IEEE Access. 2018;6:41238-41248.

[14] Liu ZS, Siu WC, Chan YL. Photo-Realistic Image Super-Resolution via Variational Autoencoders. IEEE Trans Circuits Syst Video Technol. 2021;31:1351-1365.

[15] Zhao H, Gallo O, Frosio I, Kautz J. Loss Functions for Image Restoration With Neural Networks. IEEE Trans Comp Imaging. 2017;3:47-57.

[16] Lim B, Son S, Kim H, Nah S, Mu Lee K. Enhanced Deep Residual Networks for Single Image Super-Resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. IEEE. 2017:136-144.

[17] Yang J, Wright J, Huang TS, Ma Y. Image Super-Resolution via Sparse Representation. IEEE Trans Image Process. 2010;19:2861-2873.

[18] Dong C, Loy CC, He K, Tang X. Learning a Deep Convolutional Network for Image Super-Resolution. European Conference on Computer Vision (ECCV). ECCV. 2014:184-199.

[19] Xiong Y, Guo S, Chen J, Deng X, Sun L, et al. Improved SRGAN for Remote Sensing Image Super-Resolution Across Locations and Sensors. Remote Sensing. 2020;12:1263.

[20] Wang X, Yu K, Wu S, Gu J, Liu Y, Dong C, Qiao Y, Change Loy C. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In Proceedings of the European conference on computer vision (ECCV) workshops. 2018:63-79.

[21] Senthil A, Tomlinson L. Enhancing Super-Resolution Models: A Comparative Analysis of Real-ESRGAN, AESRGAN, and ESRGAN. In Seventh International Conference on Image Processing and Machine Vision (IPMV 2025). SPIE. 2025;13636:116-123.

[22] Zhu Z, Lei Y, Qin Y, Zhu C, Zhu Y. IRE: Improved Image Super-Resolution Based on Real-ESRGAN. IEEE Access. 2023 Mar 10;11:45334-45348.

[23] Zhang K, Liang J, Van Gool L, Timofte R. Designing a Practical Degradation Model for Deep Blind Image Super-Resolution. IEEE/CVF International Conference on Computer Vision (ICCV). 2021:4791-4800.

[24] Su J, Xu B, Yin H. A Survey of Deep Learning Approaches to Image Restoration. Neurocomputing. 2019;352:1-14.

[25] Yan Q, Gong D, Zhang Y. Document Image Super-Resolution Using Structural Similarity and Sparse Representation. J Electron Imaging. 2020;29:043007.

[26] Liu Z, Luo P, Wang X, Tang X. Deep Learning Face Attributes in the Wild. IEEE/CVF International Conference on Computer Vision (ICCV). IEEE. 2018:3730-3739.

[27] Timofte R, De Smet V, Van Gool L. A+: Adjusted Anchored Neighborhood Regression for Fast Super-Resolution. Asian Conference on Computer Vision (ACCV). 2014:111-126.