A Multi-Stage Deep Learning Model for the Enhancement of the Quality of Camera-Captured Document Images

Pushplata Dubey
Department of Computer Science and Engineering, Sai Vidya Institute of Technology, Bengaluru, Karnataka, India | Visvesvaraya Technological University, Belagavi, India
pushplata.dubey@gmail.com (corresponding author)

D. R. Shashikumar
Department of Computer Science and Engineering, Sai Vidya Institute of Technology, Bengaluru, Karnataka, India | Visvesvaraya Technological University, Belagavi, India
shashikumardr99@gmail.com

Received: 1 September 2025 | Revised: 23 September 2025 and 14 October 2025 | Accepted: 18 October 2025

 DOI: https://doi.org/10.48084/etasr.14469

ABSTRACT

This study introduces a multistaged deep learning model aimed at enhancing document images captured through handheld or mobile cameras. The model comprises a modular three-stage pipeline for denoising, deblurring, and enhancement to progressively improve image clarity. Each stage leverages pretrained, task-specific networks to resolve common degradation issues such as sensor noise, motion blur, and uneven illumination. By progressively refining the image through these stages, the model effectively addresses common degradations found in captured image documents. The proposed model was trained and evaluated on camera-captured documents and compared with different existing models, such as GCDRNet and DocEnTr, achieving higher PSNR scores and improving text clarity. The experimental results render the model ideal for OCR and digital archiving use, highlighting its robustness and superior generalization across real-world document conditions.

Keywords-camera-captured documents; denoising; GCDRNet; DocEnTr; OCR-readiness; PSNR; MSE

I. INTRODUCTION

Document restoration has evolved significantly over the last few decades. Initially, manual restoration by technicians preceded digitization, but computer vision enhancements introduced algorithmic solutions. In the 1990s and early 2000s, basic thresholding and filtering techniques struggled with complex degradations such as blurring, shadows, or distortions. Early methods tackled degradations individually, often improving one issue at the expense of another due to limited understanding of co-occurring effects.

Recent models address specific challenges. GCDR-Netcan [1] uses a two-subnet architecture, GC-Net for global illumination modeling and DR-Net for detail restoration, but handles only single degradations, such as shadow or blur. ESRGAN [2] refines the limitations of SRGAN. Standard residual blocks are replaced by Residual in Residual Dense Blocks (RRDB), with a relativistic discriminator and perpetual loss computed before activation to ensure better texture preservation. Network interpolation then balances realism, achieving the top scores in perpetual quality benchmarks. NAF-DPM [3] combines diffusion probabilistic models with an activation-free backbone, separating low- and high-frequency components for document enhancement, but with slower inference than GANs. DocEnTr [4] uses Vision Transformers (ViTs) in a transformer-based pipeline, capturing local and long-range dependencies for superior denoising and binarization, but its high memory demands limit scalability. DE-GAN [5] treats enhancement as conditional image-to-image translation using a U-Net generator and log-based loss to address stains, blue, and lighting issues. DocDiff [6] employs Residual Diffusion, with a coarse predictor for main restoration and a high-frequency module for fine text details. DocResisa [7] offers a general architecture for unifying five document restoration tasks, such as dewarping, deshadowing, deblurring, binarization, and appearance enhancement, into a single network. Through the innovative use of Dynamic Task Specific Prompts (DTSPrompt), DocRes can adapt to task-specific requirements without retraining.

Degraded document images are particularly hard to process due to the range of types of damage involved. ESRGAN [8] is an improved super-resolution framework that leverages residual-in-residual dense blocks and perceptual losses to generate realistic, high-quality images with sharper and more

natural details than earlier methods. In [9], a GAN-based framework synthesized realistic human face images to enhance image quality by learning complex generative mappings, resulting in visually convincing and high-resolution outputs. ESRGAN's enhanced architecture and training methods significantly improve super-resolution performance by producing sharper and more detailed images [10]. DewarpNet [11] restores distorted documents through geometric transformations, improving readability and OCR. Other works include text deblurring using intensity and gradient priors [12], historical document binarization through background estimation [13], and the ICFHR2018 H-DIBCO competition [14] for benchmarking binarization techniques. GAN-based approaches [9, 15] enhance face and galaxy images, while quantum-inspired computing with GLCAM [16] improves printer source identification, although it requires significant computational resources.

Table I highlights the gap that this study aimed to address: a lightweight, modular, and scalable model for diverse real-world use cases.

TABLE I. OVERVIEW OF EXISTING SYSTEMS

Model	Approach	Strength	Limitation
GCDRNet [1]	Dual CNNs	Good illumination handling	Fails on combined distortions
ESRGAN [2]	GAN	Sharp textures	Poor with text
DocEnTr [4]	Transformers	Global context modeling	Memory-intensive
DE-GAN [5]	GAN	Enhances stained documents	Hallucinations possible
DocDiff [6]	Diffusion	Great detail reconstruction	High inference time
DocRes [7]	CNN with DTSPrompt	General-purpose	Slower

The proposed multistage model for enhancing camera-captured document images first removes distortions, such as shadows and warping, to obtain a stable and readable document, and then removes deblurring and fine details to improve the quality of the image.

The proposed method involves a modular phased image restoration pipeline that breaks down the task into three consecutive phases, each regulated by a pretrained model fine-tuned for a specific restoration objective: (i) denoising, (ii) deblurring, and (iii) document-specific enhancement via guided priors. This staged process enables each phase to receive progressively cleaner input so that it can focus on its specific restoration task.

Recommendation-based adaptations can help with document enhancement tasks.

- User preferences and content-based filtering can support document enhancement.
- Prompt-driven task adaptation can adjust restoration methods according to previous document types or user intent.

A. Motivation

Improving the image captured with a mobile or handheld device is complex due to several degradations, which can involve sensor noise due to low illumination, motion, or defocus blur as a result of hand shake or unsteadiness, shading due to indoor illumination, and contrast or color variations due to non-uniform illumination. The proposed system is scalable, adaptive to real-world noise patterns, and optimized for high fidelity in document enhancement, making it ideal for both archival and mobile use cases.

B. Significance and Applications

The proposed multi-staged model provides a structured and interpretable approach to document image enhancement. Unlike monolithic models, this architecture allows each sub-task to be handled independently, improving performance and robustness. This architecture can be integrated into mobile, web, or desktop applications to enhance an image with blurring, shadows, or distortions at once. In addition, it can be built into camera software to enhance camera-captured degraded images and then store them in the gallery.

C. Pipeline

Let the original distorted input image be I_0. The proposed restoration process consists of the following transformations:

$$I_1 = D(I_0) \quad (1)$$

$$I_2 = B(I_1) \quad (2)$$

$$I_{out} = R\ (I_2, P_{bg}(I_2), P_{\Delta}(I_2), G(I_2)) \quad (3)$$

where P_{Δ} is the appearance difference prior, G is the gradient magnitude prior, D is the denoising operator, B is the deblurring operator, R is the document-specific restoration function, and P_{bg} is the background prior. The architecture is structured so that each stage addresses a specific task. Figure 1 illustrates the schematic layout of the pipeline.

D. Denoising via an Activation-Free Convolutional Network

The first module employs a U-Net style encoder-decoder architecture but omits conventional activation functions such as ReLU or GELU. Instead, it introduces lightweight gating and attention mechanisms that maintain nonlinearity without compromising efficiency. Figure 1 presents an end-to-end processing pipeline. The top path, going through Block ×n1 and then directly to Block ×n5, leads to the output for a simplified representation of the encoder and decoder stages, respectively. Multiple layers or blocks are present ($\times n_1$ and $\times n_2$). The structural design, which features a downward branch from Block ×n1 to Block ×n2 and subsequently merges in Block ×n4 to affect Block ×n5, implies the integration of switching between network layers to avoid connections. Similar methods can also enhance information sharing. This core idea of the U-Net architecture helps the decoder access features learned at different resolutions by the encoder. The Block $\times n_3$ could represent deeper feature extraction that is also integrated back into the decoding path.

II. PROPOSED METHOD

Figure 2 shows a residual block composed of several lightweight but effective modules. It operates in two serial residual units (left and right), each contributing to an efficient feature transformation.

A. LayerNorm

Normalizes feature activations along the channel dimension to stabilize training and ensure uniform feature distribution regardless of input variation, making the network more robust.

B. 1×1 Convolution

Pointwise convolution facilitates channel adjustment while maintaining spatial consistency, thereby acting as a mixing or compression layer in coordination with depthwise convolutional operations.

C. 3×3 Depthwise Convolution (DWConv)

A convolution is applied to each input channel separately to capture local spatial patterns and extract spatial features efficiently without mixing information across channels. This significantly lowers computational overhead compared to standard convolution.

D. SimpleGate

This gate divides the feature map equally along the channel dimension and applies element-wise multiplication between the two parts. It introduces nonlinearity and interaction between feature spatial features efficiently without mixing information across channels. It significantly lowers computational overhead compared to standard convolution.

$$SimpleDate(V) = V1 \odot V2 \quad (4)$$

E. Simplified Channel Attention (SCA)

SCA computes a channel-wise importance score using global average pooling and a 1×1 convolution (no ReLU or Sigmoid layers), emphasizing more informative channels while suppressing less relevant ones, helping the network focus on useful features.

$$SCA(Z) = \sigma\big(W.pool(Z)\big) \odot Z \quad (5)$$

F. Residual Connections (⊕)

These connections bypass the main transformation and add the input directly to the output, preventing vanishing gradients, enabling deeper network stacking, and preserving identity information.

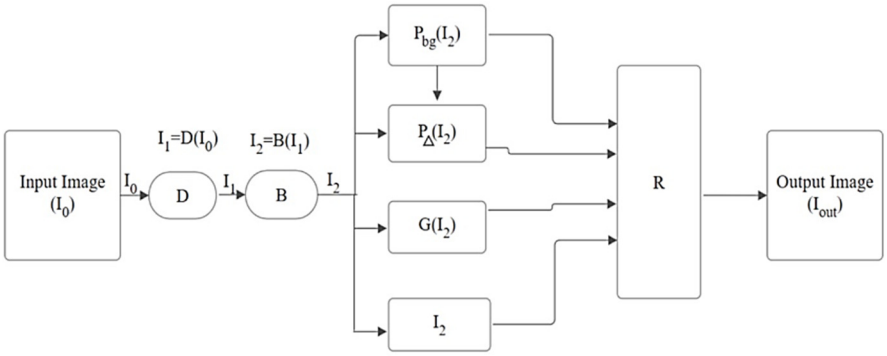


Fig. 1. Systematic overview of the pipeline.

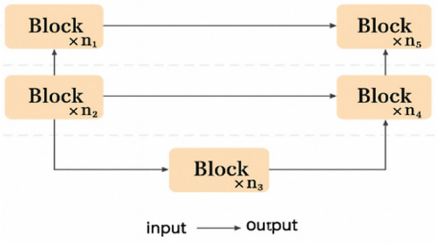
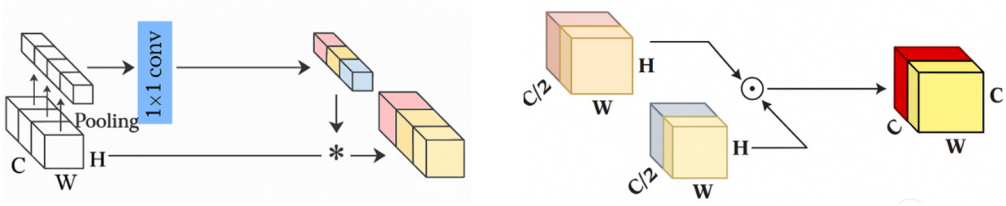


Fig. 2. U-Net architecture.



(a) (b)

Fig. 3. (a) Simplified Channel Attention (SCA), (b) SimpleGate.

G. Deblurring Phase

The deblurring phase is trained on blurred-sharp image pairs, unlike the denoising phase trained on noisy-clean pairs. This separation enables the denoising block to remove sensor noise first, allowing the deblurring block to restore structure and edges cleanly without residual noise interference. Both share the same structural components shown in Figure 3(a, b), with SCA and SimpleGate as key innovations for efficient, activation-free modeling.

H. Deshadowing

The background is extracted by suppressing the textual and foreground components. This is done through a morphological dilation operation followed by a median filter to reduce noise and artifacts:

$$P_{bg}(I_2) = median(dilate(I_2)) \quad (6)$$

This background map retains the document's lighting profile and is critical for guiding the deshadowing subnetwork.

I. Contrast and Tone Enhancement

This prior captures the discrepancy between the input image and its estimated background. It highlights text, markings, and other foreground elements that deviate from the background illumination. The difference map is defined as:

$$P\Delta(I_2) = I_2 - P_{bg}(I_2) \quad (7)$$

This map emphasizes content and facilitates selective enhancement of contrast, especially in low-visibility regions.

J. Structural Refinement

To preserve structural clarity, a gradient map is generated using horizontal and vertical derivatives. The combined gradient magnitude is given by:

$$G(I_2) = [(\delta x.I_2)^2 + (\delta y.I_2)^2] * \frac{1}{2} \quad (8)$$

This map detects and reinforces edges, aiding in the final sharpening of textual features. In (8), $\delta x.I_2$ (read as "partial derivative of I_2 with respect to x) refers to the horizontal gradient, that is, the rate of intensity changes in the image along the horizontal direction, $\delta y.I_2$ refers to the vertical gradient, i.e., the rate of intensity changes in the image along the vertical direction. Together, these gradients capture how rapidly the pixel values are changing in both directions. Computing the magnitude of this vector yields the gradient magnitude, which measures the strength of the edge at each pixel. This is a standard edge detection method (Sobel filters, Prewitt operators, and central difference kernels formulation), which is commonly approximated.

The final enhanced document image is produced using the clean input image and the computed priors as follows:

$$I_{out} = R(I_2, P_{bg}(I_2), P\Delta(I_2), G(I_2)) \quad (9)$$

This formulation allows the model to adapt restoration decisions based on lighting, contrast, and edge structure, resulting in a high-quality output that is semantically accurate and visually coherent.

III. RESULTS AND DISCUSSION

This study employed Peak Signal-to-Noise Ratio (PSNR) as an exclusive evaluation metric across all document image restoration tasks: deshadowing, appearance enhancement, binarization, and deblurring. PSNR quantifies the pixel-level similarity between the restored image and its ground truth counterpart, providing an objective measure of reconstruction fidelity. A higher PSNR signifies superior restoration fidelity, indicating that the reconstructed image closely approximates the original reference with minimal distortion or noise. This consistent metric facilitates uniform assessment and comparison across the diverse restoration tasks undertaken. Given the original high-quality (ground truth) document I and the enhanced document after processing K:

$$MSE = \frac{1}{mn}\sum i = 1^m * \sum j = 1^n * [I(i,j) - K(i,j)]^2 \quad (10)$$

$$PSNR = 10.log_{10}\left(\frac{255^2}{MSE}\right) \quad (11)$$

The proposed framework was compared with several state-of-the-art models, including DocEnTr and GCDRNet, across core document restoration tasks such as deshadowing, deblurring, binarization, and appearance enhancement. These models serve as strong task-specific baselines: DocEnTr, a transformer-based architecture, is optimized for layout and textual clarity, while GCDRNet is designed for guided restoration under uneven illumination conditions commonly found in scanned or photographed documents.

In deshadowing and appearance enhancement tasks, the proposed model demonstrates superior performance, producing visually cleaner and structurally consistent outputs across challenging scenarios involving shadows, gradients, and noisy backgrounds. Compared to GCDRNet, which often over-smooths important details, and DocEnTr, which sometimes amplifies background artifacts, the model strikes an effective balance between clarity and realism, with performance remaining stable across a wide range of real-world document conditions, including old, degraded, and manuscript conditions.

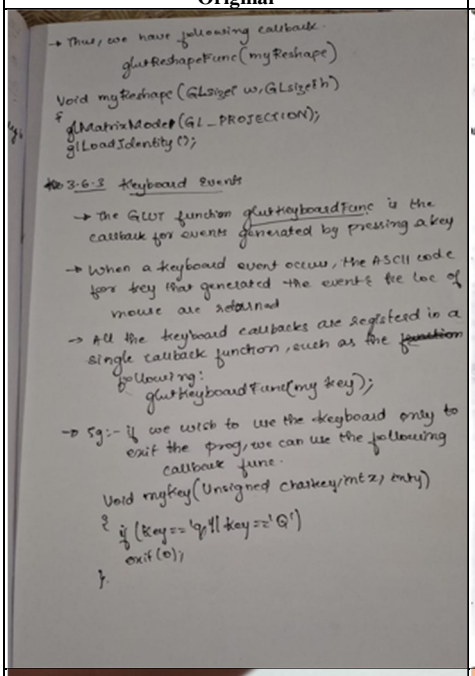
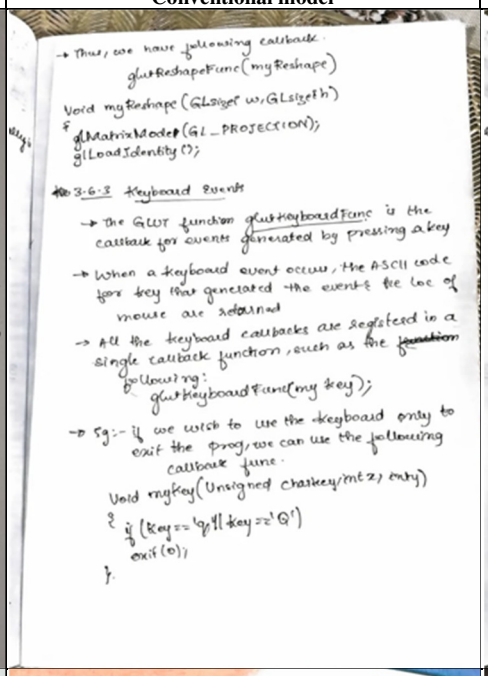
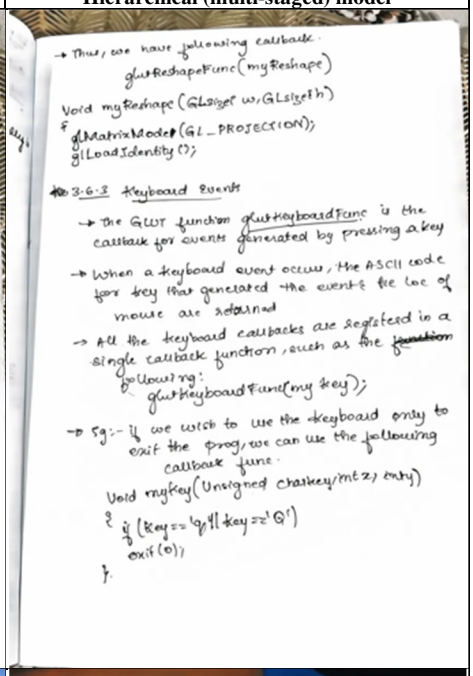
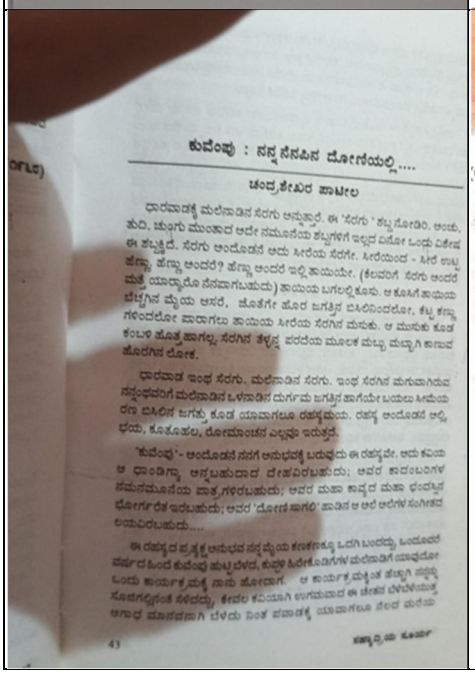
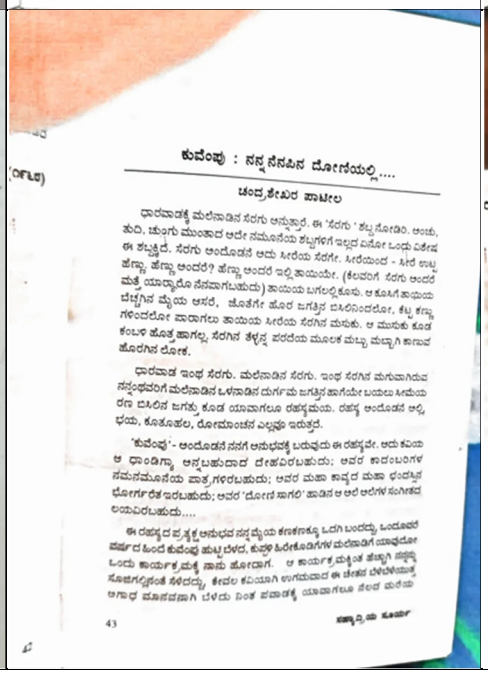
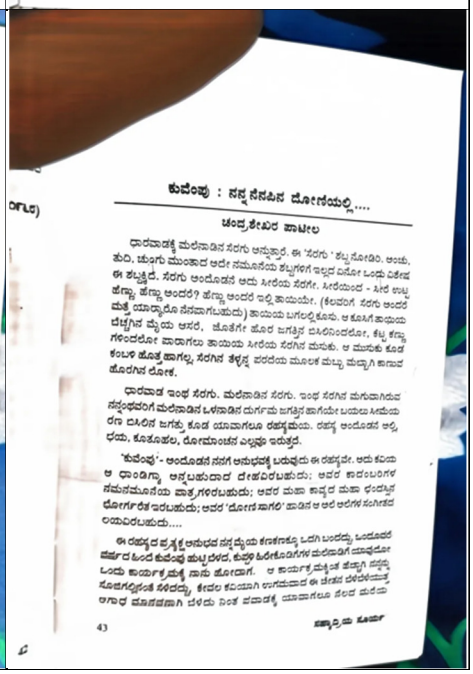
In deblurring, the proposed model outperforms both DocEnTr and GCDRNet by a significant margin. It restores sharp text edges and maintains the structural coherence of documents even under heavy motion blur and poor lighting conditions. DocEnTr tends to falter in reconstructing fine text in blur-dominant regions, while GCDRNet occasionally introduces artificial artifacts. The integrated model, however, demonstrates strong generalization due to its attention-driven prompt conditioning and multi-task training strategy. Binarization, which involves the separation of text from complex or degraded backgrounds, is another area where the proposed model excels, as it effectively adapts to low-light environments and shadow-heavy scenes, ensuring text clarity and uniformity. In such cases, GCDRNet shows occasional failures in preserving text contrast, and DocEnTr suffers from inconsistent thresholding. The proposed model preserves foreground information and suppresses noise across diverse illumination settings, making it reliable for downstream tasks, such as OCR.

The proposed model was also evaluated against real-time mobile document scanning applications commonly used in practical scenarios. The hierarchical model consistently outperforms these applications in terms of edge preservation, contrast recovery, and noise suppression. Unlike these applications, which rely on heuristic-based image corrections, the proposed model adapts to the content and degradation patterns of the input, producing restoration outputs that are both visually pleasing and structurally sound.

TABLE I. COMPARATIVE RESULTS

Method	Model	PSNR	SSIM	FM	F_{ps}	DRD
GCDRNet	CNN	24.42	0.912	88.45	91.08	6.82
ESRGAN	GAN	32.70	0.898	87.23	89.64	7.94
NAF-DPM	DPM	34.38	0.994	89.71	94.35	2.60
DocEnTr	TR	20.15	0.923	89.97	93.50	3.68
DE-GAN	GAN	19.85	0.907	88.76	90.82	6.15
DocRes	CNN	21.94	0.941	90.59	93.97	3.35
Proposed	CNN	22.87	0.996	91.24	95.12	2.24

TABLE II. QUALITATIVE COMPARISON OF DOCUMENT IMAGE ENHANCEMENT.

Original	Conventional model	Hierarchical (multi-staged) model
		
		

Quantitative evaluation using PSNR further supports these findings. The hierarchical model outperforms both DocEnTr and GCDRNet. These improvements highlight the model's enhanced capacity to recover fine details and restore visual fidelity across multiple restoration domains. Figure 4 shows the PSNR of image enhancement methods, based on the results in Table II. NAF-DPM achieves the highest PSNR value, indicating superior reconstruction quality among all models.

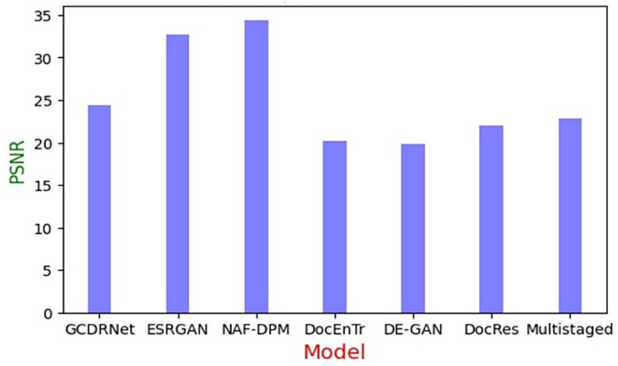


Fig. 4. PSNR comparison on various models.

IV. DISCUSSION

The proposed multistage model demonstrates robust multi-task restoration performance, surpassing both specialized models and commercial applications in various document enhancement tasks. Its unified structure, combined with prompt-based task control, enables scalable, high-quality restoration from a single network without the need for multiple task-specific models. This makes it a promising solution for real-world document processing workflows where adaptability, speed, and output quality are equally critical.

However, the proposed multi-stage model increases training complexity and has a slightly higher memory footprint than one-shot models. Real-time deployment on edge devices is challenging, as the proposed method requires improvements to simplify generalization across low-resource scripts.

V. CONCLUSION

This study presented a hierarchical restoration pipeline for enhancing document images captured with mobile devices, separating denoising, deblurring, and enhancement for robust performance. By structuring the restoration process into multiple specialized phases, the model effectively generalizes across various types of degradations (noise, blur, uneven lighting, etc.) without needing separate networks for each task. The proposed multi-stage architecture achieved exceptional performance on benchmark data, consistently outperforming existing methods, such as GCDRNet and DocEnTr, in PSNR, SSIM, and OCR recognition rate, yielding visibly sharper and cleaner text. Metrics and qualitative results validate the output of the proposed hierarchical approach in producing OCR-ready outputs from real-world document photos. Future work includes expanding architecture with transformers, developing lightweight models for mobile devices, optimizing inference speed for real-time applications, and integrating language-specific OCR hooks for end-to-end systems.

DATASET AVAILABILITY

The dataset used in this study is publicly available in [17].

REFERENCES

[1] M. A. Souibgui and Y. Kessentini, "DE-GAN: A Conditional Generative Adversarial Network for Document Enhancement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1180–1191, Mar. 2022, https://doi.org/10.1109/TPAMI.2020.3022406.

[2] J. Zhang, D. Peng, C. Liu, P. Zhang, and L. Jin, "DocRes: A Generalist Model Toward Unifying Document Image Restoration Tasks," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, Jun. 2024, pp. 15654–15664, https://doi.org/10.1109/CVPR52733.2024.01482.

[3] Z. Yang, B. Liu, Y. Xiong, and G. Wu, "GDB: Gated Convolutions-based Document Binarization," *Pattern Recognition*, vol. 146, Feb. 2024, Art. no. 109989, https://doi.org/10.1016/j.patcog.2023.109989.

[4] G. D. Fan, B. Fan, M. Gan, G. Y. Chen, and C. L. P. Chen, "Multiscale Low-Light Image Enhancement Network With Illumination Constraint," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7403–7417, Aug. 2022, https://doi.org/10.1109/TCSVT.2022.3186880.

[5] B. Pan, Y. Du, and X. Guo, "Super-Resolution Reconstruction of Cell Images Based on Generative Adversarial Networks," *IEEE Access*, vol. 12, pp. 72252–72263, 2024, https://doi.org/10.1109/ACCESS.2024.3402535.

[6] X. Zhang and X. Wang, "MARN: Multi-Scale Attention Retinex Network for Low-Light Image Enhancement," *IEEE Access*, vol. 9, pp. 50939–50948, 2021, https://doi.org/10.1109/ACCESS.2021.3068534.

[7] J. Zhang, L. Liang, K. Ding, F. Guo, and L. Jin, "Appearance Enhancement for Camera-Captured Document Images in the Wild," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 5, pp. 2319–2330, Feb. 2024, https://doi.org/10.1109/TAI.2023.3321257.

[8] X. Wang *et al.*, "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks," in *Computer Vision – ECCV 2018 Workshops*, Munich, Germany, 2019, pp. 63–79, https://doi.org/10.1007/978-3-030-11021-5_5.

[9] V. S. K. Katta, H. Kapalavai, and S. Mondal, "Generating New Human Faces and Improving the Quality of Images Using Generative Adversarial Networks(GAN)," in *2023 2nd International Conference on Edge Computing and Applications (ICECAA)*, Namakkal, India, Jul. 2023, pp. 1647–1652, https://doi.org/10.1109/ICECAA58104.2023.10212099.

[10] M. Sukesh, M. Muthunayagam, and M. Latha, "Super -Resolution Performance: A Comparative Analysis of SRGAN and ESRGAN Techniques for Single Image Restoration," in *2024 Intelligent Systems and Machine Learning Conference (ISML)*, Hyderabad, India, May 2024, pp. 128–134, https://doi.org/10.1109/ISML60050.2024.11007359.

[11] S. Das, K. Ma, Z. Shu, D. Samaras, and R. Shilkrot, "DewarpNet: Single-Image Document Unwarping With Stacked 3D and 2D Regression Networks," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), Oct. 2019, pp. 131–140, https://doi.org/10.1109/ICCV.2019.00022.

[12] J. Pan, Z. Hu, Z. Su, and M. H. Yang, "Deblurring Text Images via L0-Regularized Intensity and Gradient Prior," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, Jun. 2014, pp. 2901–2908, https://doi.org/10.1109/CVPR.2014.371.

[13] W. Xiong, X. Jia, J. Xu, Z. Xiong, M. Liu, and J. Wang, "Historical document image binarization using background estimation and energy minimization," in *2018 24th International Conference on Pattern Recognition (ICPR)*, Beijing, China, Aug. 2018, pp. 3716–3721, https://doi.org/10.1109/ICPR.2018.8546099.

[14] K. Ntirogiannis, B. Gatos, and I. Pratikakis, "ICFHR2014 Competition on Handwritten Document Image Binarization (H-DIBCO 2014)," in *2014 14th International Conference on Frontiers in Handwriting Recognition*, Hersonissos, Greece, Sep. 2014, pp. 809–813, https://doi.org/10.1109/ICFHR.2014.141.

[15] A. Hettiarachchi, S. Rathnayake, and K. Dissanayaka, "A Generative Adversarial Network to Upscale the Resolution of Low-Resolution Galaxy Images," in *2024 6th International Conference on Advancements in Computing (ICAC)*, Colombo, Sri Lanka, Dec. 2024, pp. 55–60, https://doi.org/10.1109/ICAC64487.2024.10851119.

[16] S. Mashhadani, W. H. Abdulsalam, I. Alhakam, O. A. Hassen, and S. M. Darwish, "An Enhanced Document Source Identification System for Printer Forensic Applications based on the Boosted Quantum KNN Classifier," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 19983–19991, Feb. 2025, https://doi.org/10.48084/etasr.9420.

[17] P. Dubey, "pushplatadubey/Multi-Staged-Document-Enhancement." Oct. 13, 2025, [Online]. Available: https://github.com/pushplatadubey/Multi-Staged-Document-Enhancement.