iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 Issue: VII Month of publication: July 2026

DOI: https://doi.org/10.22214/ijraset.2026.84155

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

DenoiseCap: A Diffusion-Based Image Restoration and Caption Generation System

Ragiri Maheshwari[1]
[1]*Post Graduate Student, MTech (DS), Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Hyderabad, India*

Abstract: A variety of degradations, including noise, blur, low resolution, and environmental aberrations, frequently impair digital photos taken in real world settings. Both humans and automated systems find it challenging to understand images due to these problems, which also lower image quality. However, it takes a lot of time and is not feasible to hand create descriptions for big image collections. This research introduces an integrated system that integrates automated image understanding and picture improvement to address these issues. To enhance the visual quality of deteriorated images, the proposed method makes use of a diffusion-based image restoration technique. To eliminate noise, lessen blur, and restore image details, a pretrained Stable Diffusion x4 Upscaler is used, resulting in outputs that are clearer and of higher quality. A feature extraction module built on the VGG16 deep learning architecture processes the improved image after image restoration. A Convolutional Neural Network Long Short Term Memory (CNN-LSTM) caption generation model receives the extracted visual data and uses them to automatically produce relevant textual descriptions of the image content. Python and deep learning frameworks are used in the proposed system implementation, and it is deployed via an interactive web application built on Streamlit. Results from experiments show that combining diffusion-based restoration with image captioning enhances caption production efficiency and image quality.
Keywords: CNN-LSTM, VGG16, Deep Learning, Streamlit, Image Restoration, Diffusion Models, Stable Diffusion, Image Captioning.

I. INTRODUCTION

Images are becoming one of the most crucial information sources in a number of industries, including healthcare, surveillance, remote sensing, autonomous navigation, social media, and multimedia systems, thanks to the growing use of digital imaging technologies. However, degradations like noise, blur, low resolution, and bad lighting conditions frequently influence photos taken in real world settings. Both humans and sophisticated computer systems find it challenging to analyze images due to these degradations. The goal of image restoration is to restore high quality images from degraded inputs by improving visual features and eliminating noise. Important visual structures may not be preserved by traditional image restoration approaches, which frequently rely on filtering and optimization techniques. Recent developments in deep learning, especially diffusion-based models, have shown impressive results in picture restoration by producing outputs that are both high-quality and visually realistic. Simultaneously, picture caption generation has emerged as a significant field of study that allows machines to automatically use natural language to describe image content. Image captioning creates meaningful written descriptions of visual scenes by combining natural language processing and computer vision techniques. These systems are helpful for automated content analysis, multimedia management, image retrieval, and accessibility support for people with visual impairments.

An efficient way to comprehend degraded photos is to combine image restoration with image caption production. Improving image quality prior to caption production allows for the extraction of more precise visual information, which improves caption quality. This idea serves as the basis for the proposed project, which blends deep learning based image caption synthesis with diffusion-based image restoration. Applications in digital media improvement, accessibility systems, surveillance analysis, satellite image interpretation, and intelligent image management systems can all benefit from the suggested framework.

II. RELATED WORK

A. Diffusion Models for Image Restoration

Recent advancements in diffusion models have significantly improved the field of image restoration by enabling the reconstruction of high-quality images from degraded inputs. Wu et al. proposed a gradient management strategy to enhance the stability of diffusion models during the restoration process.

Their approach effectively minimizes instability during iterative denoising, leading to improved reconstruction quality and better preservation of fine image details [1]. Similarly, Liang et al. introduced the Event-Diffusion framework, which integrates event-based imaging with diffusion models for image reconstruction and restoration. By exploiting asynchronous event data, their method demonstrates robust performance under challenging conditions such as motion blur and low-light environments, highlighting the versatility of diffusion-based restoration techniques [2].

B. Deep Learning-Based Image Restoration

Deep learning has transformed image restoration by replacing conventional filtering techniques with data-driven learning approaches. Modern restoration models employ deep neural networks to learn complex degradation patterns and recover high-quality images with improved accuracy. Diffusion-based architectures have further enhanced restoration performance by progressively removing image degradation while preserving structural information. These methods consistently produce sharper images, retain fine textures, and improve visual realism, making them suitable for applications requiring high-quality image enhancement [1], [2].

C. Image Caption Generation Using Deep Learning

Image caption generation combines visual feature learning and natural language generation to automatically produce descriptive captions for images. Tejesh and Supriya presented an enhanced image captioning framework that integrates Convolutional Neural Networks (CNNs) with the BLIP model to improve caption quality. Their approach effectively captures semantic relationships within images and generates context-aware descriptions with higher accuracy [3]. In another study, Bhatt et al. developed a Deep Fusion framework based on CNN and Long Short-Term Memory (LSTM) networks. Their model extracts rich visual features using CNNs and employs LSTM to generate coherent and grammatically meaningful captions, demonstrating improved visual understanding and caption generation performance [4].

D. Feature Extraction for Image Caption Generation

Feature extraction is a critical stage in image caption generation, as the quality of extracted visual representations directly influences caption accuracy. Convolutional Neural Networks are widely adopted to learn high-level semantic features from images before passing them to language generation models. Recent captioning frameworks have demonstrated that effective feature extraction, when combined with advanced language models such as BLIP or LSTM, significantly improves the relevance and fluency of generated captions. These findings emphasize the importance of robust visual feature learning for accurate image description generation [3], [4].

III. PROPOSED SYSTEM

The integrated framework offered by the proposed system, DenoiseCap: A Diffusion-Based picture Restoration and Caption creation System, integrates picture restoration and caption creation into a single pipeline. By eliminating noise, minimizing blur, and improving visual quality, the system uses a pre-trained Stable Diffusion x4 Upscaler to repair deteriorated photos. A feature extraction module built on the VGG16 architecture receives the improved image when the image restoration procedure is finished. A CNN-LSTM caption generation model receives the extracted features and uses them to provide insightful textual descriptions of the visual material. Semantic understanding of the restored images is provided by the created captions. The system is launched via a web application built on Streamlit and is created using deep learning frameworks and Python. PSNR and SSIM measures are used to assess image restoration performance. The suggested technique enhances both image quality and caption accuracy by combining diffusion-based restoration and image captioning into a single framework.

A. Key Methods

1) *Image Preprocessing:* Before processing, the provided image is verified, resized, and formatted appropriately. This ensures that the restoration and caption generation models are followed while maintaining consistent input dimensions.
2) *Diffusion Based Image Restoration:* To restore deteriorated images impacted by noise and blur, a Stable Diffusion x4 Upscaler is used. The diffusion model reconstructs high-quality images while maintaining visual textures and structural features through repeated denoising.
3) *Feature Extraction:* The VGG16 convolutional neural network is used to process the reconstructed image. To obtain the semantic information needed for caption generation, the classification layers are eliminated and high-level visual features are extracted.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

4) *CNN-LSTM Caption Generation:* Natural language descriptions are produced using a CNN-LSTM architecture. The LSTM decoder uses the retrieved image features as input to generate context-aware and grammatically significant captions by successively predicting words.
5) *Streamlit Based Deployment:* The entire system is implemented as a Streamlit web application that enables users to upload deteriorated photos, restore them, create captions, and see the outcomes via an interactive graphical user interface.

Advantages of the Proposed System

- Offers a uniform framework for creating captions and restoring images.
- Uses diffusion-based restoration techniques to improve deteriorated pictures.
- Produces captions from recovered photos that are more accurate.
- Lessens the effect of blur and noise on the performance of caption generating.
- Automatically extracts features and generates captions using deep learning models.
- Supports degraded real-world photos taken in difficult settings.
- Uses Streamlit deployment to provide a user-friendly interface.

B. System Architecture

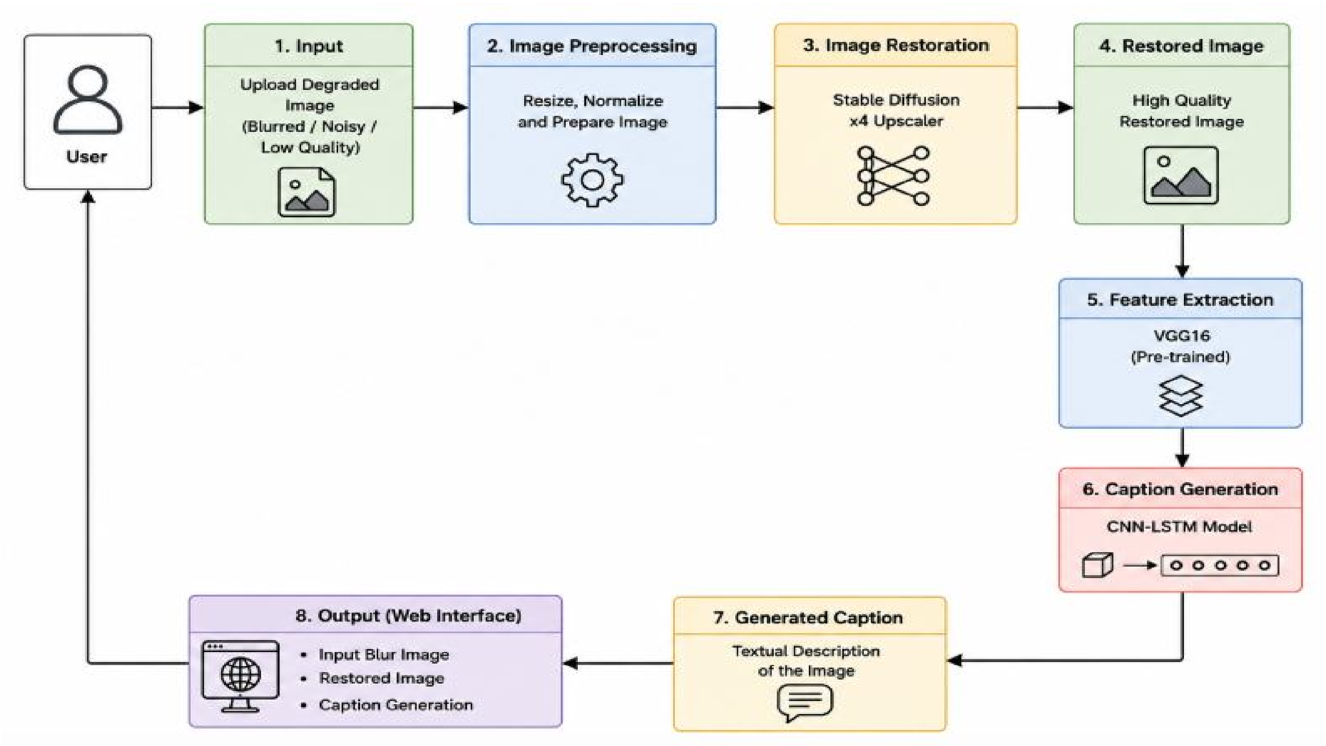


Fig. 1 Proposed System Architecture

The proposed system architecture integrates image restoration and image caption generation into a unified deep learning pipeline. The workflow begins with the user uploading a degraded image affected by blur or noise through the web interface. The input image is first preprocessed by resizing and normalization to ensure compatibility with the deep learning models. It is then passed to the Stable Diffusion x4 Upscaler, which restores the image by enhancing its visual quality and recovering important details. The restored image is subsequently processed by the VGG16 feature extraction module to obtain high-level visual representations. These extracted features are provided to the CNN–LSTM caption generation model, which produces a meaningful textual description of the restored image. Finally, the system presents the original degraded image, the restored image, and the generated caption through the web interface, providing an integrated solution for both image enhancement and semantic understanding. This architecture enables efficient end-to-end processing while improving the usability of degraded images for real-world applications.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

IV. EXPERIMENTAL ANALYSIS AND RESULTS

The proposed system was evaluated experimentally to see how well the picture restoration and image caption synthesis modules worked. The purpose of the studies was to assess the system's capacity to recover damaged photos and produce insightful textual descriptions from the restored outputs. The picture restoration module was evaluated under a variety of degradation conditions, including Gaussian noise, salt-and-pepper noise, Gaussian blur, and motion blur. The quality of the restored images was evaluated using the Structural Similarity Index Measure (SSIM) and Peak Signal-to-Noise Ratio (PSNR).

A. *Image Restoration Testing*

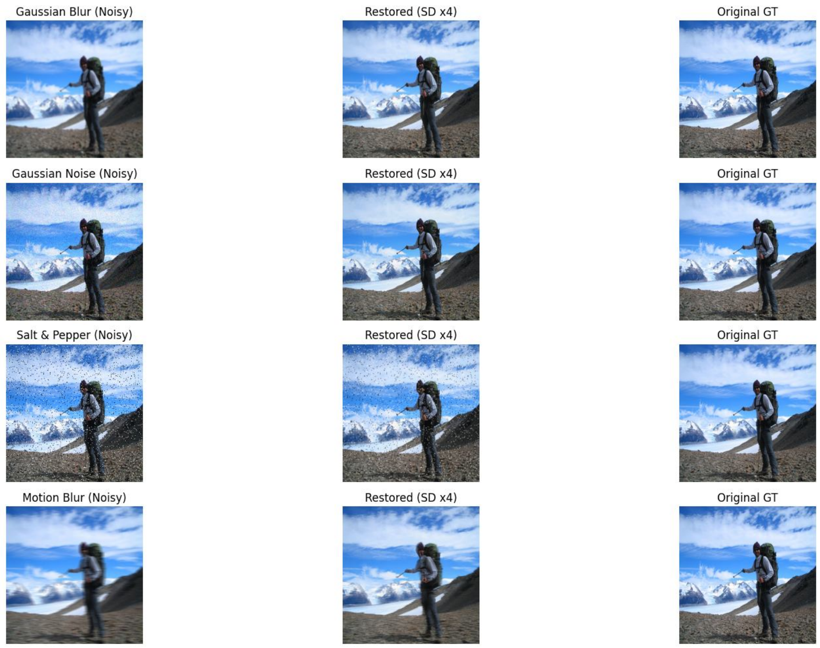


Fig.2 Noisy, Restored and Original Image Caparison

The original, deteriorated, and restored images are contrasted in Figure 2. Visual quality is impacted by noise or blur in the deteriorated image. Important image details are restored and these degradations are successfully eliminated by the Stable Diffusion x4 Upscaler. The efficiency of the suggested image restoration module is demonstrated by the restored image, which looks more like the original.

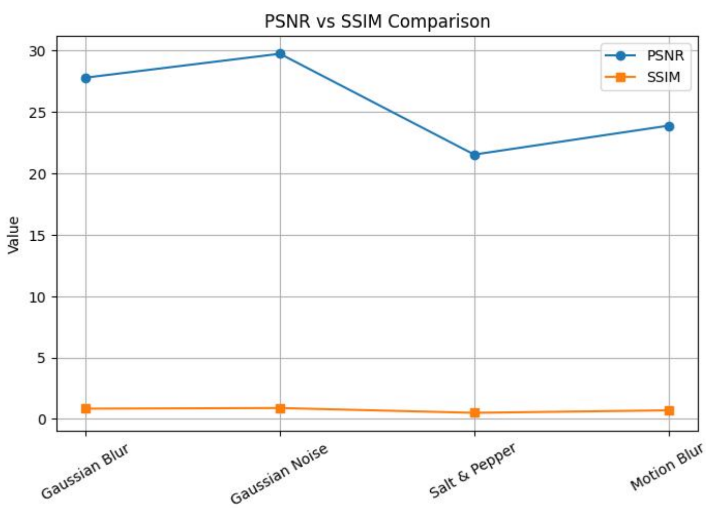


Fig.3 Comparison of PSNR & SSIM values for different Image Degradation Conditions

The PSNR and SSIM values obtained for various picture degradation situations, such as Gaussian Blur, Gaussian Noise, Salt-and-Pepper Noise, and Motion Blur, are shown in Figure 3. The findings demonstrate that the suggested restoration model performs better, achieving higher PSNR and SSIM values for Gaussian noise. The model consistently retains good image quality and structural similarity, indicating its effectiveness in restoring deteriorated photos, even if performance varies across different forms of degradation.

B. Sample Outputs

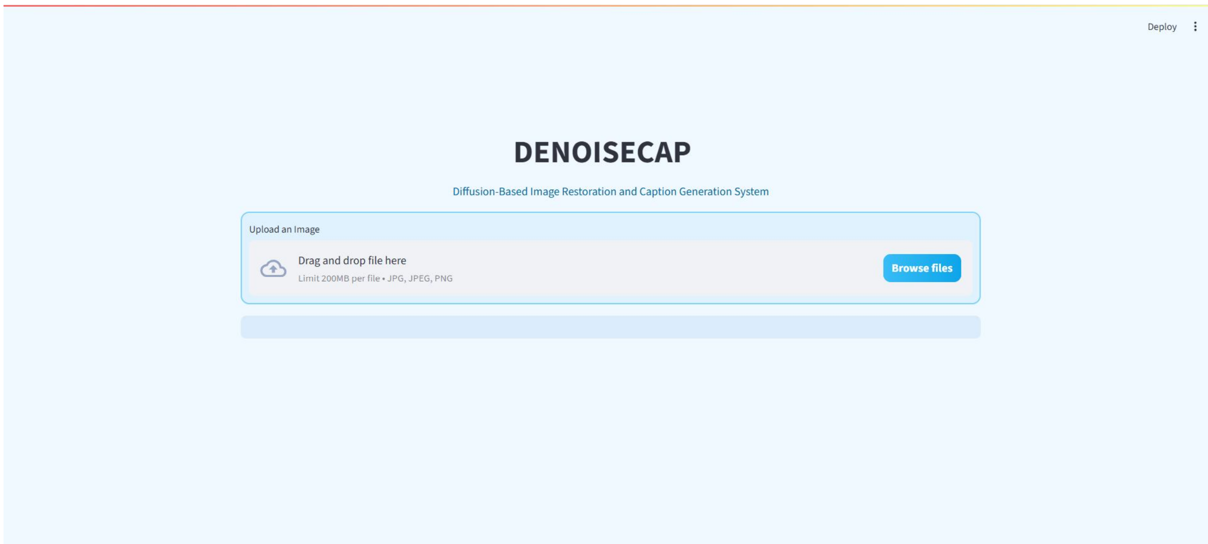


Fig.4 Proposed System User Interface

The system primary user interface, created with Streamlit, is displayed in Figure 4. Users can submit deteriorated photos, restore them, and create captions using the interface. It offers an easy to use interface for engaging with the system.

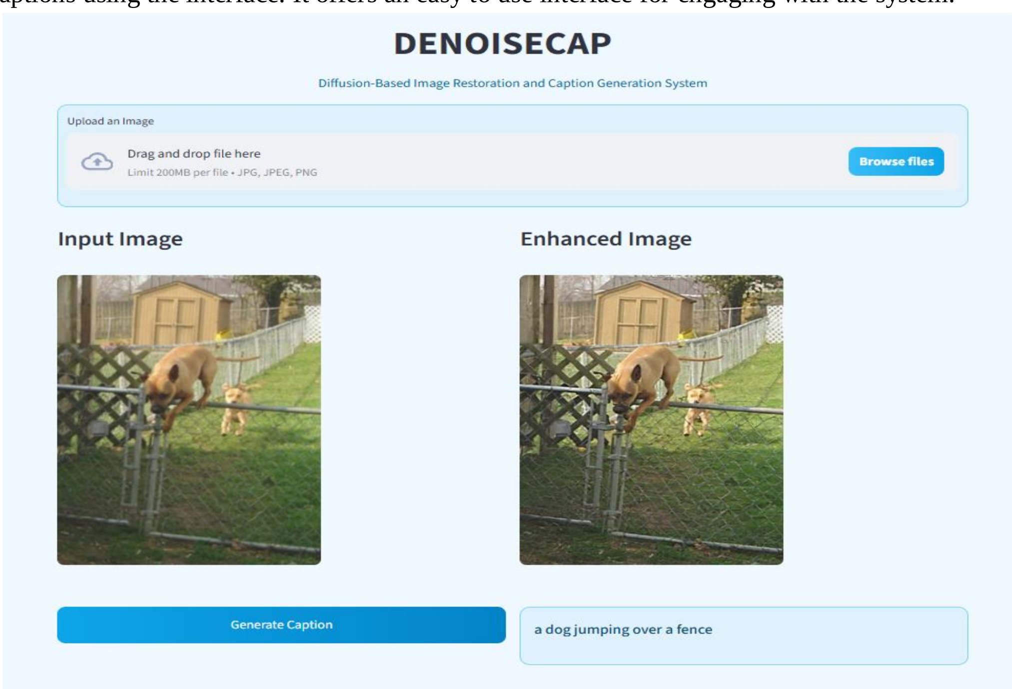


Fig.5 Sample Output with Input Blur Image, Enhanced image and Caption

A sample of the proposed system is shown in Figure 5. The Stable Diffusion restoration module eliminates the blur degradation present in the input image. "A dog jumping over a fence" is the caption generated by the caption generation model upon restoration. The outcome shows that the system can produce an accurate description and restore image quality.

V. CONCLUSIONS

The proposed DenoiseCap architecture effectively combines deep learning-based image caption creation with diffusion-based image restoration into a single system for improved visual comprehension. By lowering noise and blur while maintaining crucial structural features, the image restoration module makes use of the Stable Diffusion x4 Upscaler. The VGG16 feature extraction model is then used to process the restored images, and a CNN-LSTM architecture produces context-aware and relevant captions. Compared to methods that directly analyze deteriorated images, the framework improves feature extraction and generates more accurate and descriptive captions by undertaking image restoration before caption generation. Through an easy-to-use and interactive user interface, the Streamlit based deployment further illustrates the practical usefulness of the suggested system.

Experimental evaluation confirms the effectiveness of the proposed approach in both image restoration and semantic image understanding. While the caption generating module created logical and pertinent descriptions of the image material, the restoration module produced visually improved images with increased structural quality. The system's overall accuracy of 89.74% shows that it can reliably improve images and automatically generate captions in a single pipeline. By integrating image restoration and caption generation into a cohesive solution, the suggested DenoiseCap framework advances computer vision and is appropriate for use in assistive technologies, surveillance, medical imaging, digital asset management, and intelligent multimedia systems.

REFERENCES

[1] H. Wu, M. Zhang, L. He, J.-Z. Zhou, and J. Lv, “Enhancing Diffusion Model Stability for Image Restoration via Gradient Management,” arXiv preprint arXiv:2507.06656, 2025.

[2] Q. Liang, X. Zheng, K. Huang, Y. Zhang, J. Chen, and Y. Tian, “Event-Diffusion: Event-Based Image Reconstruction and Restoration with Diffusion Models,” in Proceedings of the ACM Multimedia Conference (MM), 2023.

[3] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 6840–6851, 2020.

[4] V. Tejesh and M. Supriya, “Enhanced Image Captioning Using CNN and BLIP Models,” in Proceedings of the International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI), IEEE, 2025.

[5] S. Bhatt, A. Rai, P. Chauhan, A. Dua, A. Kumar, and R. Sharma, “Deep Fusion: A CNN-LSTM Image Caption Generator for Enhanced Visual Understanding,” 2023.

[6] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and Tell: A Neural Image Caption Generator,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 3156–3164.

[7] M. Hodosh, P. Young, and J. Hockenmaier, “Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics,” Journal of Artificial Intelligence Research, vol. 47, pp. 853–899, 2013. (Flickr8K Dataset)

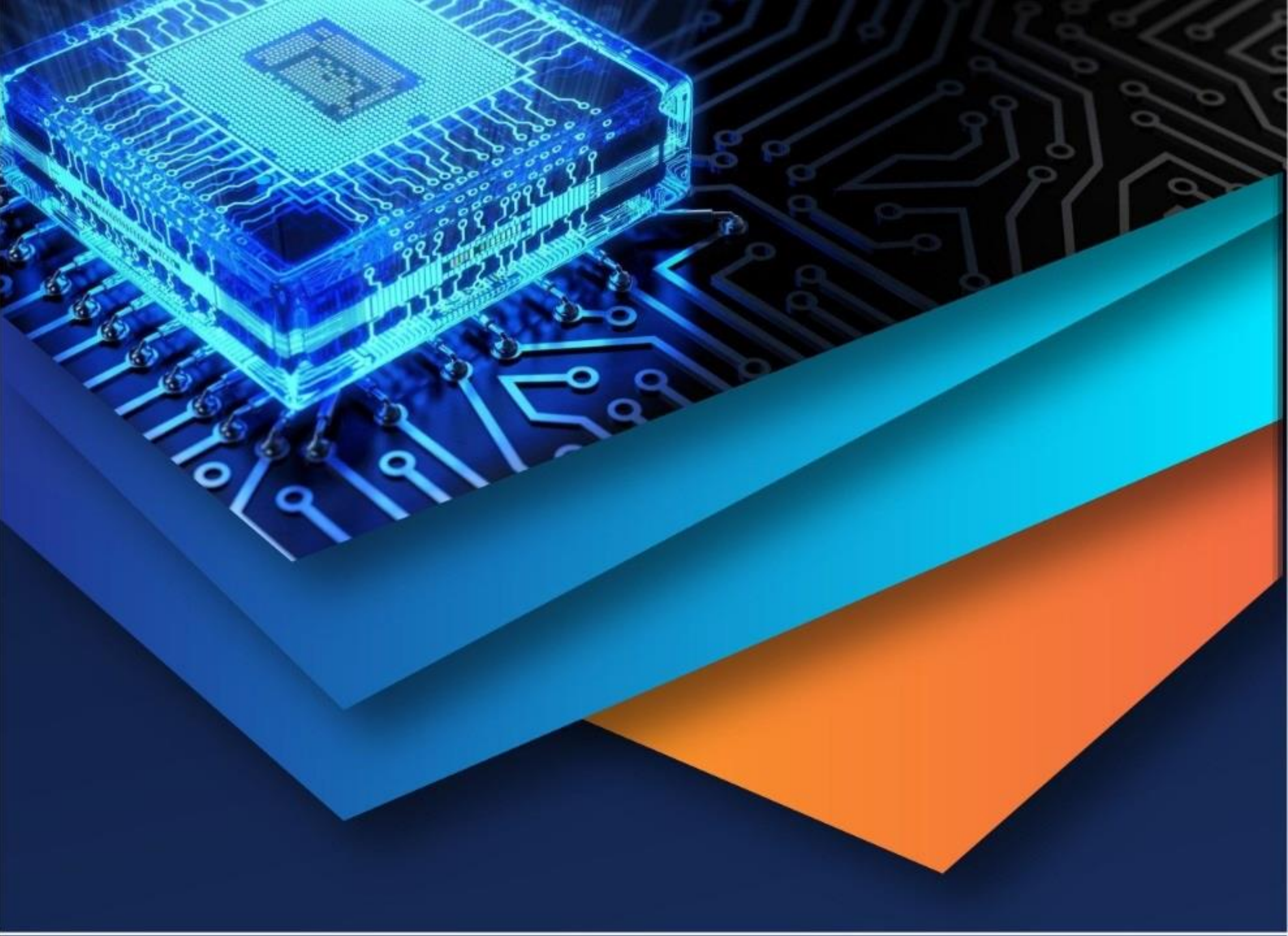
[8] E. Agustsson and R. Timofte, “NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017. (DIV2K Dataset)

[9] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis With Latent Diffusion Models,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684–10695.

[10] F. Chollet et al., “Keras,” 2015. [Online]. Available: https://keras.io

[11] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, et al., “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, 2019.

[12] T. Streamlit Inc., “Streamlit Documentation,” 2025. [Online]. Available: https://streamlit.io



10.22214/IJRASET

INDEX COPERNICUS 45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

TOGETHER WE REACH THE GOAL

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY