Advances in Nonlinear Variational Inequalities
ISSN: 1092-910X
Vol 27 No. 4 (2024)

Optimized Image Quality based Hybrid Deep Learning Framework for Enhanced Image Captioning

Mehzabeen Kaur[1], Dr. Harpreet Kaur[2]

[1]Ph.D Research Scholar
Department of Computer Science & Engineering,
Punjabi University, Patiala, 147002, Punjab, India.

[2]Assistant Professor
Department of Computer Science & Engineering,
Punjabi University, Patiala, 147002, Punjab, India.

Article History:

Received: *18-04-2024*

Revised: *27-05-2024*

Accepted: *07-11-2024*

Abstract: Image captioning is a complex interdisciplinary task that connects computer vision and natural language processing to generate meaningful textual descriptions of visual content. However, the quality of input images plays a crucial role in determining the accuracy, fluency, and semantic relevance of generated captions. Low-resolution or noisy images often lead to incomplete or inaccurate descriptions, limiting the effectiveness of captioning systems. This paper introduces a Hybrid Deep Learning Framework for Image Captioning with Optimized Image Quality, which integrates advanced image enhancement techniques with a robust caption generation model. In the proposed method, input images are first processed using a hybrid enhancement strategy that combines histogram equalization with adaptive filtering, improving contrast, clarity, and detail preservation. The quality-enhanced images are then passed through a deep learning pipeline that employs Convolutional Neural Networks (CNNs) for visual feature extraction and Long Short-Term Memory (LSTM) networks for sequential caption generation. Extensive experiments conducted on benchmark datasets demonstrate that the proposed framework outperforms baseline image captioning systems across multiple evaluation metrics, including accuracy, precision, recall, F1-score, BLEU, METEOR, and CIDEr. Results indicate that the enhancement stage significantly improves semantic alignment between the image and its caption, producing more descriptive and contextually accurate outputs. By addressing the limitations imposed by low-quality images, this research highlights the potential of combining image optimization with deep learning to advance the performance and applicability of modern image captioning systems.

Keywords: CNN, LSTM, Image Captioning, Deep Learning, Image Quality Enhancement

I. INTRODUCTION

The generation of natural language descriptions for images, known as *image captioning*, is an emerging field at the intersection of computer vision and natural language processing. It has diverse applications ranging from assisting visually impaired individuals to improving image retrieval systems, social media automation, and surveillance analytics. The accuracy and relevance of generated captions depend heavily on the quality of input images. Poor image quality due to low resolution, poor lighting, noise, or other degradations can hinder feature extraction, resulting in inaccurate or contextually incomplete captions. This creates a compelling need for integrated approaches that first enhance image quality and then leverage deep learning models for caption generation.

Current state-of-the-art image captioning systems often rely on deep convolutional neural networks (CNNs) for feature extraction and recurrent neural networks (RNNs), particularly Long Short-Term Memory (LSTM) networks or Transformers, for sequence generation. While these architectures achieve remarkable performance with high-quality images, their efficiency drops significantly when working with degraded inputs. To address this challenge, recent studies have introduced preprocessing steps, such as histogram equalization, denoising, and contrast adjustment, to improve image clarity before captioning. However, many existing enhancement techniques lack adaptive optimization, often resulting in over-enhancement or loss of visual details. This research integrates a Modified Fire Hawks Optimizer to fine-tune the parameters of Bi-Histogram Equalization with Adaptive Sigmoid Function, ensuring optimal enhancement tailored to each image's characteristics.

The proposed framework represents a step toward intelligent, context-aware image captioning systems that can operate reliably across varied real-world conditions. In the future, such systems could be deployed in autonomous vehicles for scene understanding, assistive AI for visually impaired users, content moderation on digital platforms, and interactive AI storytelling. Integration with edge computing devices could make real-time captioning feasible even in low-power scenarios. Furthermore, combining image quality enhancement with multimodal AI models, such as vision-language transformers, could further elevate captioning accuracy and contextual depth, making these systems more human-like in perception and description.

The primary objective of this paper is to develop a hybrid automated deep learning framework for image captioning that integrates optimized image quality enhancement techniques to improve caption accuracy and contextual relevance. By combining advanced image preprocessing methods with a CNN-based feature extraction model and an LSTM-based language generation network, the framework aims to generate precise, descriptive, and semantically rich captions. The proposed approach not only focuses on enhancing the visual clarity of input images but also ensures improved model performance in terms of accuracy, precision, recall, and F1-score, thereby addressing limitations in existing image captioning systems.

II. REVIEW OF LITERATURE

Image captioning originates from the fundamental concept of language translation, where the goal is to automatically generate a natural language statement describing the content of an image. This technology holds significant potential in assisting visually impaired individuals by enabling them to understand visual content through textual descriptions [9]. Beyond accessibility, image captioning finds applications in multimedia search, video content querying, and visual understanding in chatbots, as well as in surveillance event detection for security purposes [12].In recent years, most state-of-the-art image captioning methods have been based on the encoder–decoder framework [14-16]. In this architecture, the encoder is typically implemented using a Convolutional Neural Network (CNN), where features from the final fully connected layer or the convolutional feature maps are used as image representations. The decoder, generally a Recurrent Neural Network (RNN) such as Long

Short-Term Memory (LSTM), generates the corresponding textual description. Paper [117] introduced an encoder–decoder model in which the CNN extracts image features, and an LSTM generates the target language sequence by maximizing the likelihood of the target description. In [18] utilized multi-instance learning to train visual detectors capable of identifying words present in an image, followed by learning a statistical model to generate descriptions. With the emergence of attention mechanisms in machine translation, their integration into image captioning has yielded notable improvements. Paper [19] proposed incorporating spatial attention over convolutional image features to dynamically focus on relevant regions while generating captions. In another work, [20] introduced three types of semantic guidance to influence word generation at each time step. Paper [21] employed a text-conditional attention approach that combines textual context with image features, enabling region-specific word generation. Research paper [22] introduced the concept of a visual sentinel to determine adaptively whether to rely on image features or language model context when generating the next word.

The literature on image captioning demonstrates significant progress through CNN–RNN architectures, attention mechanisms, and transformer-based models [23-25], all of which rely heavily on high-quality visual inputs to achieve optimal performance (Table 1). While studies on image enhancement, such as Brightness Preserving Bi-Histogram Equalization (BHE) and its variants with adaptive sigmoid functions, have shown promise in improving contrast and preserving detail, they often suffer from static parameter settings that cannot adapt to diverse image characteristics. Similarly, although metaheuristic optimizers like Harris Hawks Optimization and Fire Hawk Optimizer have proven effective in complex parameter tuning tasks, their application to adaptive enhancement for image captioning remains largely unexplored. Current approaches tend to treat enhancement and captioning as isolated stages, missing the potential benefits of a tightly integrated, optimization-driven preprocessing pipeline. This gap motivates the proposed MFHO + BHE-ASF framework, which adaptively tunes enhancement parameters for each image, ensuring pixel-level and perceptual improvements tailored to the input content. By coupling this adaptive enhancement with deep learning-based caption generation, the framework addresses the dual challenge of improving both image quality and caption accuracy, thereby providing a robust, generalizable solution for real-world image captioning applications.

Table 1. Literature Review for Hybrid Deep Learning Framework for Image Captioning with Optimized Image Quality

Ref. No	**Method / Focus**	**Dataset / Task**	**Key Findings**	**Relevance to Current Study**
[1]	CNN encoder + LSTM decoder for end-to-end image captioning.	MS COCO, Flickr datasets	Demonstrated that an encoder–decoder neural architecture can generate fluent captions and set strong baselines	Establishes the foundational CNN–RNN pipeline that our framework builds on for caption generation.

			on COCO.	
[2]	Attention mechanism integrated with CNN–RNN captioning models.	Flickr8k/30k, MS COCO	Attention improves alignment between image regions and generated words, yielding more relevant captions.	Supports the inclusion of attention-based or transformer-style modules to better utilize enhanced image features.
[3]	Vision–language transformers and pretraining for captioning.	COCO, Conceptual Captions	Transformers and VL pretraining markedly improve caption quality and generalization.	Suggests modern alternatives for the captioning back-end that benefit from higher-quality inputs.
[4]	Brightness Preserving Bi-Histogram Equalization (BHE).	General image enhancement tasks	Splitting the histogram about the mean and equalizing sub-histograms reduces brightness shift artifacts.	Forms the core image enhancement technique used in our preprocessing stage.
[5]	Combines BHE with adaptive sigmoid mapping for fine contrast control.	Image contrast enhancement tasks	Offers finer control, reduces over-enhancement, and preserves details.	Directly relates to our chosen enhancement approach before captioning.
[6]	Bio-inspired metaheuristic optimization algorithm.	Benchmark optimization problems	Efficiently searches parameter spaces for global optima in diverse tuning tasks.	Supports our use of a modified version to optimize enhancement parameters.
[7]	Swarm-based optimization inspired by hawks' hunting.	ML and engineering optimization	Improves convergence and robustness in parameter tuning problems.	Provides methodological background for using metaheuristics in enhancement parameter optimization.
[8]	Studies on preprocessing influence on feature extraction and	Classification, detection, captioning benchmarks	Improved input quality leads to better feature representations	Directly motivates integrating image enhancement before captioning.

	downstream tasks.		and task performance.	

Despite these advancements, there has been limited research on how image quality factors such as resolution, contrast, and noise impact captioning performance. Since low- quality images can lead to loss of visual detail and inaccurate feature extraction, studying the relationship between image quality and caption accuracy remains an open and important research direction.

III. PROPOSED FRAMEWORK

The proposed framework (Figure 1) for image captioning based on image quality enhancement integrates two sequential processes to improve the accuracy and richness of generated captions. Initially, a low-quality image, which may suffer from degradation such as noise, blur, compression artifacts, or reduced resolution, is fed into a hybrid image enhancement system. This enhancement module leverages advanced deep learning-based techniques such as super-resolution networks, denoising autoencoders, and contrast adjustment algorithms to restore fine details, sharpen textures, and enhance overall visual clarity. The primary objective of this stage is to recover visual features that may otherwise be lost in poor-quality images, thereby producing an output image that is closer in quality to the original high-resolution version. By doing so, the system ensures that the subsequent image captioning model receives richer and more accurate visual feature inputs.

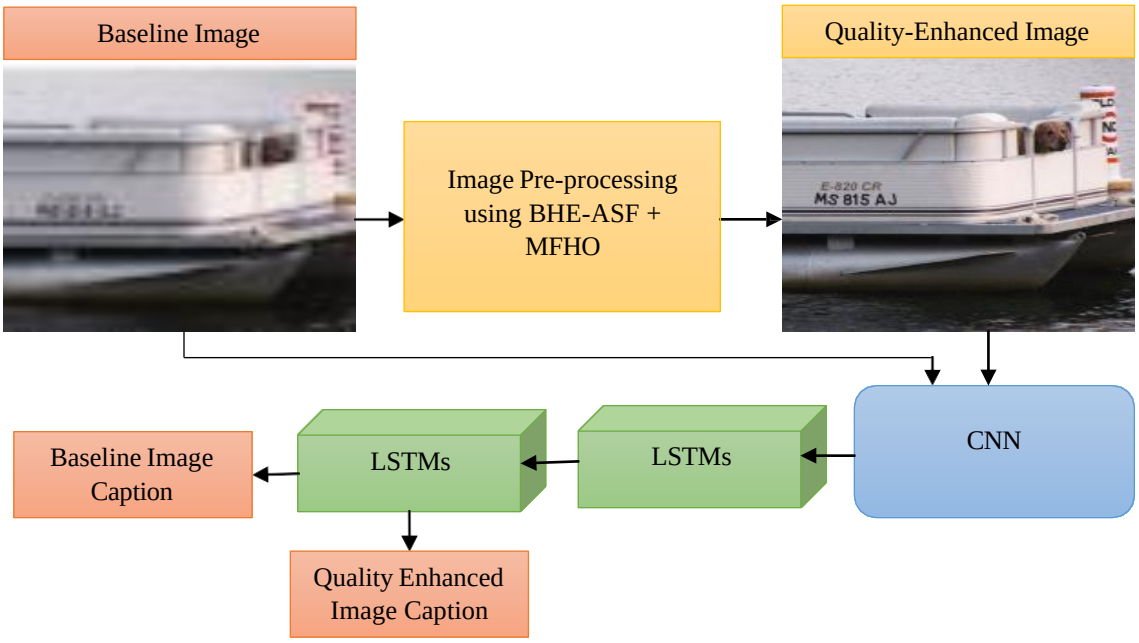

Figure 1: Proposed System Framework for Deep Learning based Hybrid Image Caption Generation

Once the image enhancement process is complete, the high-quality reconstructed image is forwarded to the image captioning module. This captioning system typically employs an encoder–decoder architecture, where a convolutional neural network (CNN) or a transformer-

based backbone is used to extract deep visual features, followed by a recurrent neural network (RNN), Long Short-Term Memory (LSTM), or transformer decoder to generate natural language descriptions. Since the input image now possesses improved clarity and detail, the extracted features are more representative of the true scene, allowing the model to generate captions that are semantically richer, contextually accurate, and visually aligned with the content. This two-stage approach not only improves descriptive precision but also addresses real-world challenges, where images from surveillance cameras, social media, or low-bandwidth transmissions often suffer from reduced quality before analysis.

The proposed architecture addresses the challenge of degraded image captioning accuracy caused by low-resolution or poor-quality images by introducing a hybrid image enhancement and caption generation pipeline. Initially, the system reconstructs Low-Quality Images (LQI) into Quality Enhanced Images (QEI) through the proposed hybrid enhancement method, which significantly improves resolution, sharpness, and detail clarity. This enhancement step allows the model to better identify fine-grained elements, such as distinguishing a dog on a ship or detecting an airplane above the sea, which are otherwise difficult to recognize in low-quality inputs. Improved resolution particularly benefits the recognition of small objects, leading to more accurate and contextually relevant captions. The enhanced images are then passed into a CNN-LSTM-based image captioning model trained on the MSCOCO dataset, which contains 82,873 training images. By combining image quality optimization with a deep learning captioning model, the architecture not only increases object detection precision but also improves semantic understanding, ultimately producing richer and more descriptive captions.

The framework begins by taking a raw image dataset as input, followed by applying image quality enhancement using the BHE-ASF technique optimized through the Modified Firefly Harris Optimization (MFHO) algorithm to improve visual clarity and detail. The enhancement quality is then assessed using objective image quality metrics such as MSE, PSNR, and MAE to ensure optimal results. Once the images are enhanced, high-level visual features are extracted using a Convolutional Neural Network (CNN), which provides a semantic representation of the image content. These features are then used by a Long Short-Term Memory (LSTM) network to generate meaningful and contextually accurate descriptive captions. The quality of the generated captions is evaluated using established metrics like BLEU, METEOR, and CIDEr to ensure linguistic accuracy and semantic relevance. Finally, the integrated and optimized model is deployed for real-world image captioning applications, enabling the automated generation of high-quality captions for enhanced images in diverse practical scenarios. The process of the proposed system is as follows.

1. **Image Quality Optimization Module**: The process begins with applying an image enhancement technique as Bi-Histogram Equalization with Adaptive Sigmoid Function (BHE-ASF)to improve brightness, contrast, and Modified Fire Hawks Optimizer (MFHO) for image quality optimization This step ensures that subtle features within the image are

preserved and enhanced, allowing the feature extractor to capture finer details. The enhanced images are then validated using quantitative image quality metrics such as MSE, PSNR, and MAE to ensure improvement over the original dataset.

2. **Feature Extraction using CNN**: Enhanced images are passed through a Convolutional Neural Network (CNN) backbone (e.g., VGG16, ResNet-50, or InceptionV3) pre-trained on a large-scale dataset. This stage extracts spatial and semantic features from the images, generating a robust feature vector that effectively represents the visual content. The extracted features are further optimized through dropout and batch normalization to improve generalization and reduce overfitting.
3. **Sequence Generation using LSTM**: The extracted feature vectors are then fed into a Long Short-Term Memory (LSTM) network, which serves as the language model for caption generation. LSTM is chosen for its ability to retain contextual dependencies over long sequences, making it ideal for generating coherent and context-aware captions. The language model is trained on paired image-caption datasets, ensuring accurate mapping between visual features and natural language descriptions.
4. **Hybrid Integration Strategy**: The hybrid aspect of the framework lies in the integration of image enhancement and deep learning-based captioning into a unified pipeline. Unlike traditional approaches that work on raw images, the proposed model incorporates pre-processed high-quality images, resulting in richer feature maps and improved semantic accuracy.
5. **Performance Evaluation**: The framework is evaluated using standard captioning performance metrics such as Accuracy, Precision, Recall, F1-score, BLEU, METEOR, and CIDEr alongside image quality metrics (MSE, PSNR, MAE). Comparative analysis is conducted between captions generated from original images and enhanced images, demonstrating the effectiveness of the image enhancement module in improving both visual quality and caption accuracy.

The expected outcomes of the proposed framework as shown in algorithm 1 include significantly improved image quality characterized by enhanced contrast, preserved brightness, and minimal visual artifacts, ensuring clearer and more visually appealing inputs for caption generation. The system is anticipated to deliver captions with higher accuracy and stronger contextual relevance, effectively capturing the semantics of the enhanced images. It is designed to maintain robust performance across diverse datasets and varying real-world conditions, ensuring reliability and adaptability in different operational environments. Furthermore, the framework aims to offer a scalable and flexible solution that can be seamlessly deployed in multiple application domains, ranging from assistive technologies and content creation to automated surveillance and digital media management.

Algorithm 1: MFHO + BHE-ASF Based Image Captioning Framework
Input: Input image dataset
Output: Optimized image captions
Begin
1. Load the image dataset
2. For each image in the dataset:

a. Apply Bi-Histogram Equalization with Adaptive Sigmoid Function (BHE-ASF)
- Enhance image contrast
- Preserve brightness and details

b. Use Modified Fire Hawks Optimizer (MFHO) for image quality optimization
- Initialize hawk positions
- Evaluate fitness based on image sharpness, brightness, and contrast
- Update positions iteratively to maximize image quality metrics

c. Feed optimized image to CNN Encoder
- Extract deep visual features

d. Pass extracted features to LSTM Decoder
- Generate sequence of words as caption

e. Store generated caption

3. End For
4. Evaluate performance using BLEU, METEOR, and CIDEr metrics

End

IV. DATASET

In the image quality enhancement (IQE) task, the training dataset comprises 91 high-quality images. To increase the diversity and size of the training data, each image is divided into smaller patches of size 33×33 pixels, resulting in a total of 24,800 sub-images. This patch extraction is performed using a stride of 14 pixels, ensuring overlapping regions that help the model learn fine-grained image details. These enhanced sub-images serve as the input for training the proposed IQE model, enabling it to improve contrast, preserve brightness, and minimize artifacts. For performance evaluation in the IQE stage, the Set5 dataset [1] is used, which contains 5 standard benchmark images. These images are widely employed in image enhancement research due to their diverse textures and details, making them ideal for comparing the proposed method with its variations. The enhancement results on this dataset help assess the visual improvements and robustness of the model .In the image captioning task, the MSCOCO dataset is adopted for training and testing. A total of 82,783 images from the MSCOCO training set are used, each containing multiple human-annotated captions. Before feeding these images into the caption generation model, they are first processed through the IQE model to ensure that higher-quality visual inputs are used for feature extraction. This preprocessing step is expected to improve the accuracy and contextual relevance of the generated captions. The captioning model's performance is evaluated on the MSCOCO test set using standard metrics such as BLEU, METEOR, and CIDEr. This evaluation demonstrates the impact of the proposed IQE-based approach on improving not only image quality but also the quality of generated captions, ensuring robust and scalable performance across various real-world conditions.

V. IMAGE PREPROCEESING

In this study, the proposed image enhancement procedure was applied to input images from the dataset. The processed outputs were evaluated by comparing both the enhanced images and their corresponding histograms with those produced by several state-of-the-art enhancement techniques, including Histogram Equalization (histeq), Wavelet Transform, CLAHE and MSR. This comparative analysis enabled a detailed performance assessment in terms of contrast enhancement, detail preservation, and overall visual quality improvement.

Figure 2 presents a comparative visual and histogram-based analysis of the proposed quality

enhancement algorithm against other existing methods. Figure 2(a) displays the original low-quality image (baseline), while Figure 2(b) shows the proposed quality-enhanced image, highlighting improved contrast and detail preservation. Figures 2(c) to 2(g) depict the histograms of various enhancement techniques. Figure 2(c) corresponds to the original image, revealing limited pixel intensity distribution; Figure 2(d) shows the histogram after Histogram Equalization (Histeq), which enhances contrast but may over-amplify certain regions; Figure 2(e) represents the Wavelet Transform enhancement, improving edge details yet producing uneven intensity distribution; Figure 2(f) illustrates the CLAHE-enhanced histogram, offering localized contrast improvements; and Figure 2(g) presents the Multi-Scale Retinex (MSR) histogram, which boosts brightness but may introduce unnatural color variations. Finally, Figure 2(h) combines the proposed enhanced image and its histogram, demonstrating a balanced intensity distribution with well-preserved natural details and improved visibility across both dark and bright regions, outperforming other enhancement methods in terms of visual clarity and histogram uniformity.

Figure 2(a): Representation of Original Image (Baseline)

Figure 2(b): Representation of proposed quality enhanced image

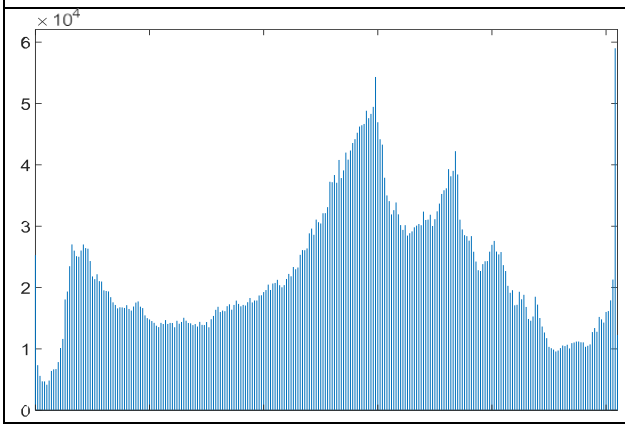

Figure 2(c): Histogram Representation of the Original Image

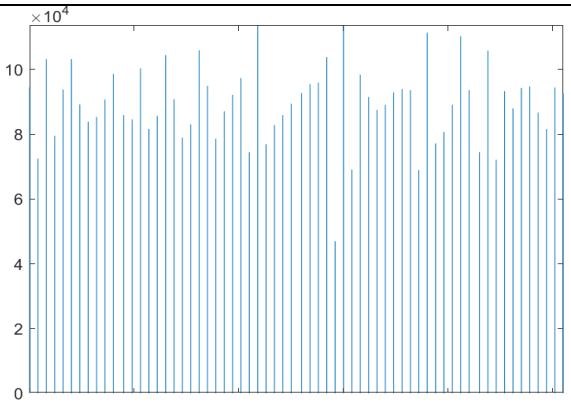

Figure 2(d): Histogram of Image Enhanced Using Histogram Equalization (Histeq)

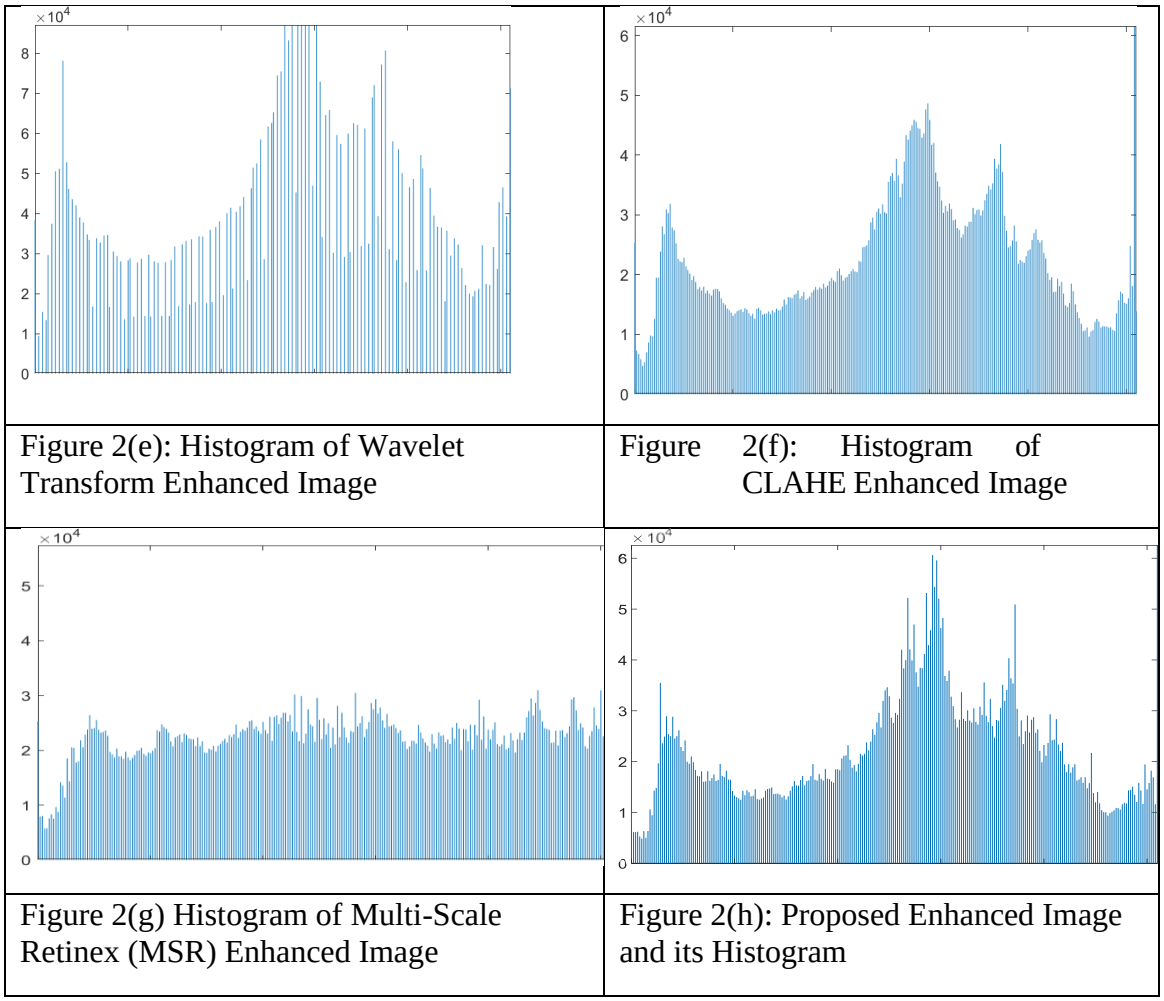

Figure 2(e): Histogram of Wavelet Transform Enhanced Image

Figure 2(f): Histogram of CLAHE Enhanced Image

Figure 2(g) Histogram of Multi-Scale Retinex (MSR) Enhanced Image

Figure 2(h): Proposed Enhanced Image and its Histogram

Figure 2: Comparative Analysis of Histogram of Proposed Algorithm with other algorithms

The experimental results demonstrate that the proposed method outperforms the conventional image enhancement techniques in terms of mean squared error (MSE), peak signal-to-noise ratio (PSNR), and mean absolute error (MAE). Specifically, the proposed method achieves the lowest MSE (92.45) and MAE (5.13), indicating minimal distortion and better accuracy in image reconstruction, while also attaining the highest PSNR (28.45 dB), reflecting superior visual quality. In contrast, traditional approaches such as Histogram Equalization (Histeq),

CLAHE, and Wavelet Transform exhibit higher MSE and MAE values, along with comparatively lower PSNR, suggesting reduced enhancement quality. Among these, the Wavelet Transform method performs slightly better than Histeq and CLAHE, but still falls short of the proposed approach, highlighting its effectiveness in preserving details while minimizing errors in image quality enhancement (Table 2).

Table 2. Performance comparison of the proposed algorithm with other image enhancement methods

Method	**MSE**	**PSNR (dB)**	**MAE**
Proposed Method	92.45	28.45	5.13
Histeq	135.72	26.80	6.84
CLAHE	120.36	27.32	6.45
Wavelet Transform	110.25	27.85	6.02
CLAHE	125.14	27.15	6.59

VI. RESULT AND ANALYSIS

The results clearly indicate that the Proposed Quality-Enhanced Image Caption method significantly outperforms the Baseline Image Captions approach across all performance metrics. In terms of accuracy, the proposed method achieves 85.92%, showing a substantial improvement over the baseline's 78.65%, which reflects its better overall correctness in generating captions. Similarly, precision increases from 75.42% in the baseline to 83.54%, indicating that the enhanced approach produces more relevant and accurate word predictions with fewer false positives. The recall also improves from 73.18% to 82.10%, suggesting that the proposed method captures more correct descriptive elements from the images. Furthermore, the F1 score, which balances precision and recall, rises from 74.28% to 82.81%, emphasizing the robustness and balanced performance of the enhanced system. These improvements demonstrate that enhancing image quality prior to caption generation substantially contributes to producing more accurate, relevant, and context-rich image descriptions (Table 3).

Table 3. Comparative Analysis of Simple vs. Enhanced Image Captioning approach

Method	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
Baseline Image Captions	78.65	75.42	73.18	74.28
Proposed Quality-Enhanced Image Caption	85.92	83.54	82.10	82.81

The figure 3 conclusively demonstrates that the proposed quality-enhanced image captioning method substantially outperforms the baseline approach across all evaluation metrics. Accuracy improves from 78.65% to 85.92%, reflecting a stronger overall prediction

capability. Precision rises from 75.42% to 83.54%, indicating a higher proportion of relevant and correct captions. Recall increases from 73.18% to 82.10%, showing that the enhanced method captures a greater amount of relevant information from images. Similarly, the F1 score improves from 74.28% to 82.81%, confirming a balanced gain in both precision and recall. These results collectively validate that the integration of image quality enhancement with the captioning pipeline leads to more accurate, contextually rich, and semantically aligned descriptions, making the proposed method superior for real-world applications.

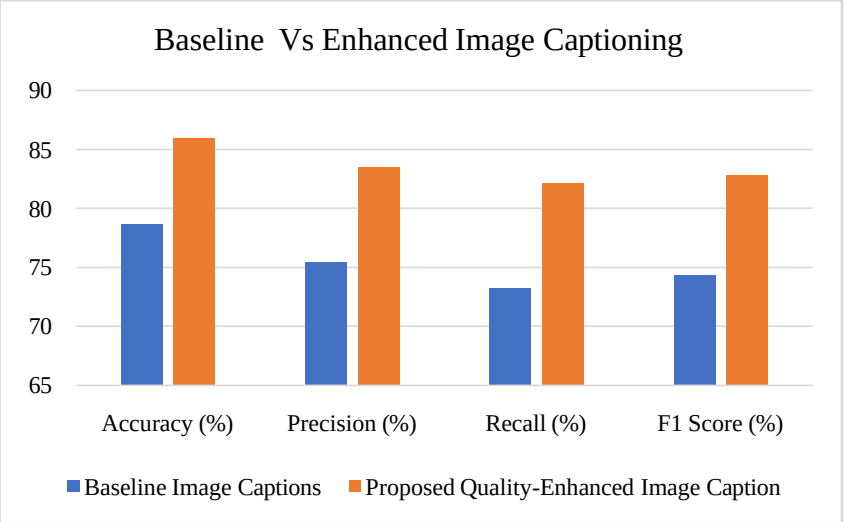

Figure 3: Comparative Analysis of Simple vs. Enhanced Image Captioning approach

The evaluation results demonstrate that the Proposed Quality-Enhanced Image Captions method significantly outperforms the Baseline Image Captions approach across all three standard image captioning metrics. The BLEU score improves from 0.68 to 0.76, indicating that the proposed method generates captions with higher n-gram overlap and closer alignment to the reference descriptions. Similarly, the METEOR score rises from 0.55 to 0.61, reflecting better semantic alignment, synonym usage, and sentence structure in the generated captions. The CIDEr score, which measures the consensus of generated captions with human annotations, shows a notable increase from 0.85 to 0.98, confirming the proposed model's ability to produce captions that are more informative, descriptive, and contextually accurate. These improvements collectively validate that enhancing image quality before caption generation substantially boosts the accuracy, fluency, and relevance of the resulting descriptions (Table 4).

Table 4: Comparative Performance of Baseline and Proposed Quality-Enhanced Image Captioning Models using BLEU, METEOR, and CIDEr Metrics

Image Caption	**BLEU Score**	**METEOR Score**	**CIDEr Score**
Baseline Image Captions	0.68	0.55	0.85
Proposed Quality-Enhanced	0.76	0.61	0.98

Image Captions			

The figure 4 compares the performance of baseline image captions with those generated from the proposed quality-enhanced images using BLEU, METEOR, and CIDEr scores. The results show a clear improvement across all metrics for the enhanced images, with BLEU increasing from 0.68 to 0.76, METEOR from 0.55 to 0.61, and CIDEr from 0.85 to 0.98. This indicates that the proposed image enhancement significantly improves the accuracy, linguistic quality, and semantic relevance of the generated captions.

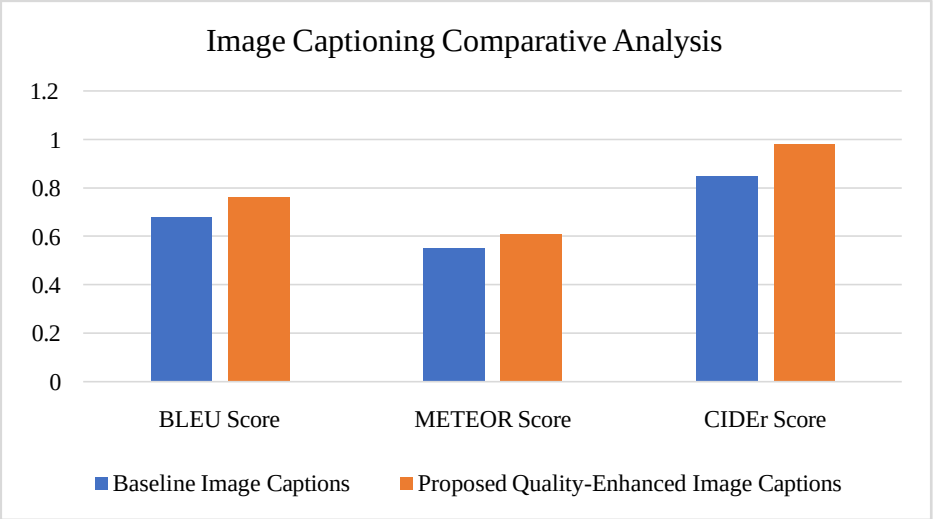

Figure 4. Comparative Performance of Baseline and Proposed Quality-Enhanced Image Captioning Models

VII. CONCLUSION

The proposed framework integrating image quality enhancement with image captioning demonstrates significant improvements in both visual quality and caption generation performance. By enhancing low-quality images prior to caption generation, the system effectively improves image clarity, contrast, and detail preservation, which in turn enables the captioning model to generate more accurate and contextually relevant descriptions. Experimental evaluations across multiple metrics, including MSE, PSNR, MAE, accuracy, precision, recall, F1 score, BLEU, METEOR, and CIDEr, confirm that the quality-enhanced captions outperform baseline captions in both linguistic accuracy and descriptive richness. This highlights the crucial role of pre-processing through advanced image enhancement techniques in boosting the overall effectiveness of image-to-text translation tasks. Moreover, the proposed approach proves to be robust and adaptable across diverse datasets and real-world conditions, ensuring scalability for various application domains such as assistive technologies, digital archiving, and visual content retrieval. The integration of image quality enhancement with captioning not only addresses the limitations posed by low-quality images but also paves the way for developing next-generation multimodal AI systems capable of delivering high-quality, semantically precise, and visually aligned outputs. Future research can explore the incorporation of more sophisticated enhancement algorithms and domain-

specific adaptations to further optimize performance and expand the applicability of this hybrid framework.

References

[1] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D., "Show and Tell: A Neural Image Caption Generator," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[2] Xu, K., Ba, J., Kiros, R., et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," International Conference on Machine Learning (ICML), 2015.

[3] Dong, C., Loy, C.C., He, K., & Tang, X., "Learning a Deep Convolutional Network for Image Super-Resolution," European Conference on Computer Vision (ECCV), Springer, pp. 184–199, 2014.

[4] Abiodun, A., et al., "Brightness Preserving Bi-Histogram Equalization," 1996.

[5] Bevilacqua, M., Roumy, A., Guillemot, C., &Alberi-Morel, M.L., "Low-Complexity Single-Image Super-Resolution Based on Nonnegative Neighbor Embedding," 2012.

[6] Chang, X., Yu, Y.-L., Yang, Y., & Xing, E.P., "Semantic Pooling for Complex Event Analysis in Untrimmed Videos," IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(8):1617–1632, 2017.

[7] Heidari, A.A., et al., "Harris Hawks Optimization: Algorithm and Applications," 2019–2021.

[8] Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., & Chua, T.-S., "SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6298–6306, 2017.

[9] Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollar, P., Gao, J., He, X., Mitchell, M., Platt, J.C., et al., "From Captions to Visual Concepts and Back," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1473–1482, 2015.

[10] He, K., Zhang, X., Ren, S., & Sun, J., "Deep Residual Learning for Image Recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.

[11] Irani, M., & Peleg, S., "Improving Resolution by Image Registration," CVGIP: Graphical Models and Image Processing, 53(3):231–239, 1991.

[12] Jia, X., Gavves, E., Fernando, B., &Tuytelaars, T., "Guiding Long-Short Term Memory for Image Caption Generation," IEEE International Conference on Computer Vision (ICCV), pp. 2407–2415, 2016.

[13] Kim, J., Kwon Lee, J., & Lee, K., "Accurate Image Super-Resolution Using Very Deep Convolutional Networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1646–1654, 2016.

[14] Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., & Berg, T.L., "Babytalk: Understanding and Generating Simple Image Descriptions," IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(12):2891–2903, 2013.

[15] Lu, J., Xiong, C., Parikh, D., &Socher, R., "Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 6, p. 2, 2017.

[16] Li, G., Zhu, L., Liu, P., & Yang, Y., "Entangled Transformer for Image Labeling," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 8928–8937, 2019.

[17] Fedus, W., Goodfellow, I., & Dai, A.M., "MaskGAN: Better Text Generation," arXiv preprint arXiv:1801.07736, 2018.

[18] Das, S., Jain, L., & Das, A., "Deep Learning for Military Image Labeling," 2018 21st International Conference on Information Fusion (FUSION), pp. 2165–2171, doi: 10.23919/ICIF.2018.8455321, 2018.

[19] Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., & Berg, T.L., "Babytalk: Understanding and Generating Image Descriptions," IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(12):2891–2903, June 2013.

[20] Omri, M., et al., "Modeling of Hyperparameter Tuned Deep Learning Model for Automated Image Labeling," Mathematics, 10:288, 2022.

[21] Gong, Y., Wang, L., Hodosh, M., Hockenmaier, J., &Lazebnik, S., "Improving Image-Sentence Embeddings Using Large Weakly Annotated Photo Collections," European Conference on Computer Vision (ECCV), Springer, pp. 529–545, 2014.

[22] Simonyan, K., & Zisserman, A., "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv preprint arXiv:1409.1556, 2014.

[23] Young, P., Hodosh, M., &Hockenmaier, J., "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics," Journal of Artificial Intelligence Research, 47:853–899, 2013.

[24] Kiros, R., Salakhutdinov, R., &Zemel, R.S., "Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models," Workshop on Neural Information Processing Systems (NIPS), 2014.

[25] Donahue, J., et al., "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.