ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

A study on temporal consistency and comparative performance of a NASNet-based deep learning model for endoscopic video denoising

Amulya H. Gopalakrishna*, Gopalakrishna Kanathur

Department of Electronics and Communication Engineering, Faculty of Engineering and Technology, Jain (Deemed-to-be University), Bengaluru, Karnataka, India

Abstract

Image processing plays an important role in endoscopic diagnosis; noise introduced by low illumination, sensor limitations, motion artifacts, and endoscope movement within the human body can significantly reduce visual clarity. Image and video quality affect the visibility of fine structures, making diagnosis difficult. Deep learning-based methods have shown promising results, but many existing models struggle to maintain temporal consistency across consecutive frames. This paper presents an enhanced video denoising framework based on a neural architecture search network (NASNet) model with a channel attention mechanism. The proposed method was optimized using a genetically modified jellyfish search optimization (GM-JFSO) algorithm. The proposed framework was evaluated using the publicly available Hyper–Kvasir dataset. It aims to preserve spatial image quality and temporal consistency across consecutive video frames by effectively reducing noise while retaining structural details. Comparative evaluation against NASNet, GoogLeNet, DenseNet, and ResNet showed that the proposed method outperformed them, achieving a peak signal-to-noise ratio of 49.07 dB, a structural similarity index measure (SSIM) of 0.938, a mean squared error of 0.584, and a temporal SSIM of 0.8313 across the evaluated sequence. The GM-JFSO algorithm improves the proposed model's performance through hyperparameter optimization, while the channel attention mechanism enhances feature representation and noise suppression. The experiment shows the effectiveness of integrating the optimization technique with the deep learning model for video denoising and highlights the importance of evaluating temporal consistency in medical imaging applications.

Keywords: Video denoising; Neural architecture search network; Channel attention; Genetically modified jellyfish search optimization; Hyper–Kvasir; Temporal consistency; Endoscopic video

***Corresponding author:** Amulya H. Gopalakrishna (amulyagopalakrishna97@gmail.com)

Citation: Gopalakrishna AH, Kanathur G. A study on temporal consistency and comparative performance of a NASNet-based deep learning model for endoscopic video denoising. *Artif Intell Health*. doi: 10.36922/AIH026340125

Received: August 20, 2026

Revised: September 25, 2026

Accepted: October 16, 2026

Published online: November 17. 2026

Copyright: © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Image processing plays an important role in modern healthcare by helping in precise diagnosis and clinical decision-making. Among imaging systems, endoscopy is widely used for real-time visualization of internal organs such as the gastrointestinal tract,

respiratory system, and other internal organs. To identify anatomical structures such as benign regions, mucosal patterns, and lesions, high-quality video is crucial for early detection and decision-making.[1,2] Endoscopic live video is usually viewed using clinical software such as CaptureIT Pro, which enables real-time analysis during endoscopic procedures.

Despite recent advances in medical imaging, endoscopic videos are often degraded by noise due to low-light conditions, sensor limitations, rapid scope movement, and fluids in the organs. These degradations can significantly reduce visual clarity and may obscure important anatomical structures. Therefore, effective denoising methods are required to reduce unwanted noise while preserving the important structural and textural information in endoscopic images and video frames.[3,4]

Traditional image denoising methods have been widely used to reduce noise and improve image quality. However, these methods may lose important image details when the noise level is high, or the image contains complex structures. With the development of deep learning, convolutional neural network-based approaches have been increasingly applied to image denoising and restoration. Deep learning-based methods can learn image degradation characteristics from training data and provide better denoising performance than conventional approaches.[5,6]

Attention-based architectures have further shown the ability to improve feature representation by focusing on important information within the image. Channel attention mechanisms can assign different importance to feature channels and help preserve relevant image features during denoising. Such attention mechanisms have also been investigated in medical image-processing applications.[7] The convolutional block attention module (CBAM), introduced by Woo *et al.*,[8] combines channel and spatial attention mechanisms to refine feature representations and enhance the extraction of informative features.

For video denoising, information from consecutive frames can provide additional cues to reduce noise and maintain visual continuity. Existing deep learning-based video denoising methods such as DVDNet and FastDVDnet have explored using neighboring frames to improve video restoration performance.[9,10] Other approaches have further explored multi-frame and spatio-temporal information for video restoration. These studies highlight the importance of evaluating not only individual-frame quality but also the consistency of processed frames across a video sequence.

Recent studies have also investigated deep learning-based methods for improving the quality of endoscopic images and videos. Low-light enhancement, endoscopic image restoration, and attention-based approaches have been explored to improve the visibility of important structures under challenging imaging conditions.[10-14] Publicly available endoscopic datasets have further supported the development and evaluation of deep learning methods for gastrointestinal endoscopic image and video analysis.[15]

A deep learning framework has been proposed for quality assessment and restoration in video endoscopy,[16] while low-light enhancement has been investigated for high-speed endoscopic videos.[17] Recent studies have discussed the increasing application and demand of deep learning in endoscopic image processing.[18] Other approaches have investigated endoscopic image enhancement using Retinex-based methods.[19-21] Among these datasets, Hyper–Kvasir provides a comprehensive collection of gastrointestinal endoscopic images and videos.[22]

Neural architecture search network (NASNet) provides transferable architectures for image recognition and can also serve as a backbone for feature extraction in image-processing applications.[23] Video/Image denoising studies have also investigated conventional and learning-based approaches to improve video quality.[24-26] These developments demonstrate the importance of combining effective feature extraction, denoising techniques, and temporal evaluation for video restoration applications.

Unlike existing approaches that mainly focus on spatial noise suppression, this study proposes an enhanced video denoising framework based on the NASNet model with a channel attention mechanism and genetically modified jellyfish search optimization (GM-JFSO) for endoscopic video denoising. The channel attention mechanism improves the feature representation by highlighting informative channels, while the GM-JFSO effectively optimizes hyperparameters. Beyond spatial denoising, the proposed framework explicitly considers temporal consistency across consecutive video frames to evaluate frame-to-frame stability and visual continuity. We compared the proposed framework with deep learning baseline models to assess its effectiveness.

The structure of this paper is as follows: Section 2 presents the literature review, Section 3 describes the methodology, Section 4 presents results and discussion, and Section 5 concludes the study and outlines future scope.

2. Literature review

2.1. Traditional image and video denoising

Image denoising has been widely studied using conventional filtering methods and transform-based

techniques. Traditional approaches such as spatial filtering and wavelet-based methods aim to suppress noise while preserving important structural details. Block-matching and 3D filtering (BM3D) has been widely used because it exploits similarities among frames and performs collaborative filtering in the transform domain.

Traditional denoising approaches provide an important foundation for image restoration because they remove unwanted variations while retaining image structures. Their performance depends on the characteristics and intensity of the noise in the input data. In medical imaging applications, preserving fine structures is critical because excessive smoothing can reduce the visibility of important anatomical details.

2.2. Deep learning-based image denoising

Deep learning-based techniques have improved image denoising by learning the complex mappings between noisy and clean images. Feature-attention-based methods such as RIDNet have shown the advantages of attention mechanisms in real-image denoising.[27]

Several restoration networks have been subsequently proposed. CBDNet addresses the blind denoising of real photographs by considering the realistic characteristics of noise.[28] MPRNet uses a multi-stage architecture for image restoration,[29] while MIRNet uses multi-scale feature representations for image restoration and enhancement.[30] More recently, transformer-based models such as SwinIR,[31] Restormer,[32] and Uformer[33] have shown that attention-based mechanisms can model long-range dependencies for image restoration.

Deep learning models are also an important consideration in medical image analysis. In endoscopic image analysis, the robustness evaluation has been investigated to understand the behavior of deep neural networks under different conditions.[34]

The development of deep learning-based methods shows the increasing use of learned feature representations for image restoration tasks. Attention mechanisms help a network identify informative representations, while multi-stage, multi-scale, and transformer-based methods offer different ways to extract and refine image features. Medical images can differ from general photographic datasets because they may contain specific anatomical structures, illumination conditions, and acquisition-related variations.

For applications such as endoscopy, maintaining structural information during denoising is critical because the images may contain fine tissue patterns and boundaries. A denoising framework should limit the unwanted noise while avoiding the excessive modification of the original visual structures. Proper feature extraction and refinement mechanisms can help meet the requirements of endoscopic image restoration.

2.3. Deep learning-based video denoising

Unlike image denoising, video denoising can obscure information that is available across neighboring frames. DVDNet combines spatial and temporal denoising steps and uses motion compensation to incorporate temporal information, while FastDVDNet simplifies the process by avoiding explicit optical-flow estimation and instead processes multiple neighboring frames to achieve efficient video denoising.

Other video restoration approaches, such as BasicVSR and BasicVSR++, have used feature propagation and alignment across neighboring frames for video restoration.[35,36] These studies demonstrate that explicit temporal modeling can provide additional information for video restoration when multiple neighboring frames are incorporated into the network. ClearNet has further investigated autoencoder-based denoising for endoscopy images.[37]

The Hyper–Kvasir dataset provides a large collection of gastrointestinal endoscopic images and videos, while Kvasir-SEG provides segmented gastrointestinal polyp images for computer-aided endoscopic image analysis.[38]

When multiple frames are used, they provide additional information that might help to distinguish the noise from the structures that remain consistent across the neighboring frames. Incorporating temporal information also requires the network to model relationships between frames. This can increase the computational requirements. The choice between frame-level and multi-frame processing depends on the design objectives and the requirements of the particular application.

In the current study, the proposed framework works on individual endoscopic frames by using the NASNet-based architecture and the implemented channel attention mechanism. The temporal information is not directly provided as a multiple-frame input to the network. Instead, it evaluates temporal consistency separately across consecutive denoised frames. This distinction allows the spatial denoising performance and the frame-to-frame behavior of the resulting video sequence to be examined separately to evaluate the performance.

2.4. Attention and NAS-based architectures

Attention mechanisms allow neural networks to place greater importance on informative features. The channel-wise attention method was introduced through

squeeze-and-excitation networks, and the combined channel-spatial attention method was introduced through the convolutional block attention module; both have shown effectiveness in adaptive feature recalibration. Restoration methods such as MPRNet and MIRNet have further demonstrated the benefits of effective feature extraction and refinement in image restoration applications.

Neural architecture search (NAS) methods provide an approach for designing an efficient deep-learning framework. NASNet introduced transferable architectures discovered through neural architecture search and demonstrated their effectiveness for image-recognition tasks. Such methods can serve as a backbone for feature extraction for image-processing techniques.

Attention-based feature refinement is appropriate for image denoising, as useful structural information and noise-related information can be represented differently within feature maps. Assigning different importance to feature representations allows the network to focus on information that contributes to reconstruction while reducing the influence of less important features. This motivates incorporating a channel attention mechanism into a denoising framework.

The selection of the base architecture is also important in deep learning-based image restoration. NASNet provides a neural architecture search-based design for feature extraction. In the proposed framework, NASNet is combined with an attention mechanism to evaluate its ability to represent features from noisy endoscopic frames.

Parameter optimization is also an important feature of deep learning-based model development. Different parameters, such as learning rates, filter configurations, and training parameters, can affect the reconstruction performance of a denoising model. In the proposed framework, the GM-JFSO architecture is incorporated to search the predefined parameter space using the reconstruction error as the optimization objective. The combination of NASNet-based feature extraction with attention-based feature refinement and parameter optimization forms the basis of the proposed framework.

2.5. Image quality evaluation

The performance of the image denoising algorithms is commonly evaluated using quantitative metrics such as peak signal-to-noise ratio (PSNR), mean squared error (MSE), and structural similarity index measure (SSIM). SSIM evaluates the structural similarity by considering factors such as luminance, contrast, and structural information.[39] Although these metrics provide useful quantitative measures of restoration quality, video denoising also requires consideration of the temporal behavior of consecutive denoised frames. Previous work has examined different deep learning-based approaches for video restoration and the importance of spatial and temporal information in video-processing applications.[40]

Spatial metrics provide an objective way to compare denoised output frames with the corresponding reference frame. PSNR reflects reconstruction error, whereas MSE directly represents the average squared difference between the reconstructed and reference images. SSIM provides additional structural comparison and can complement the pixel-level measures during evaluation.

3. Methodology

3.1. Overview of the proposed framework

The framework consists of four stages: extracting and preprocessing endoscopic video frames; frame-level denoising using a NASNet-based architecture with a channel attention mechanism; optimizing hyperparameters using the GM-JFSO algorithm; and evaluating the denoised frames using spatial image-quality metrics and temporal consistency analysis.

The proposed methodology takes endoscopic video as input, which is first separated into individual frames. The frames are then converted to RGB and resized to 331 × 331 pixels to match the input requirements of the NASNetLarge-based architecture. Pixel values are normalized to the range [0,1]. Noisy frames are then generated from the corresponding clean frames to form paired input-reference data for supervised denoising.

The proposed methodology uses NASNetLarge as the feature-extraction backbone and incorporates an attention mechanism to refine the extracted feature representation before reconstructing the denoised frame. GM-JFSO technique is used to search for suitable hyperparameter configurations using reconstruction error as the optimization criterion.

The denoised frames are finally evaluated using MSE, PSNR, and SSIM. Temporal consistency is evaluated between consecutive denoised frames using SSIM. The overall framework is illustrated in Figure 1.

3.2. Dataset preparation and frame preprocessing

The frames were extracted from the endoscopic video before being provided to the denoising network. The implementation loaded the corresponding clean and noisy images from separate directories and verified the image filenames before forming input–output pairs. Every valid image pair was then read using the Open Source Computer

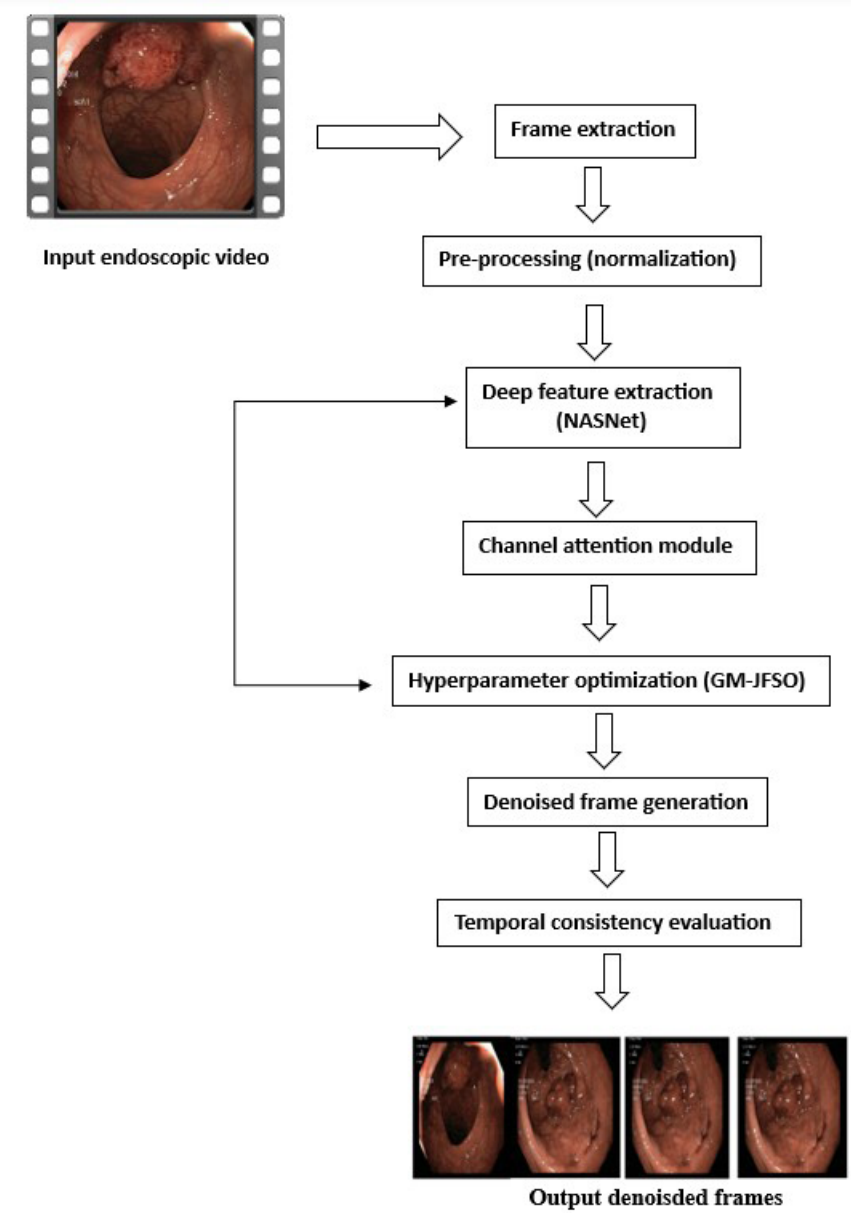


Figure 1. Proposed model for video denoising

Vision Library (OpenCV); the frames were resized to 331 × 331 pixels and converted from BGR to RGB. Normalization converted the original 8-bit pixel values to the range [0,1].

The training implementation used train_test_split with test_size = 0.1 and random_state = 42. Thus, 90% of the available frame pairs were used for training and 10% for validation.

3.3. Noise generation

Synthetic noise degradation was introduced to generate the noisy input frames while retaining the original clean frames, which were used as reference images. For experimentation and implementation, both Gaussian and salt-and-pepper noise were used. For salt-and-pepper noise degradation, the intensity parameter was taken as the percentage of pixels affected by the degradation.

For an intensity value I, the probability of an affected pixel is represented as:

$$P(\text{noise}) = \frac{I}{100} \tag{1}$$

Approximately half of the affected pixels were assigned the maximum intensity value (255), representing salt noise, while the remaining affected pixels were assigned the minimum intensity value (0), representing pepper noise. Salt-and-pepper noise with an intensity value of 10 was used for evaluation.

3.4. NASNet-based channel attention architecture

The proposed denoising network used NASNetLarge as the primary feature-extraction backbone. NASNetLarge was initialized using ImageNet-pretrained weights, and the input size was defined as 331 × 331 × 3.

The feature representation, which was obtained from the selected NASNet layers, was then provided to the implemented attention block. The attention block consisted of convolutional transformations of the feature representation, element-wise multiplication, softmax activation, feature modulation, and residual addition. A final convolutional layer with three output channels reconstructed the RGB denoised frame, which was followed by upsampling. The reconstructed output was obtained using a three-channel convolutional layer with sigmoid activation, followed by upsampling to generate the denoised RGB frame.

3.5. Genetically modified jellyfish search optimization

The proposed framework used the GM-JFSO algorithm to optimize the selected model and training parameters. The optimization process combined the jellyfish search optimization algorithm's search behavior with genetic operations to improve exploration and exploitation of the parameter search space.

The optimization method considered seven parameters, including learning rate, decoder filter sizes, convolutional filter sizes, and batch size. The learning rate was searched within the range of 1×10^{-5} to 1×10^{-3}. The decoder filter parameters were searched in the ranges 128–512, 64–256, and 32–128, while the two convolutional filter parameters were searched in the ranges 32–128 and 16–64. Batch size was also included as a parameter in the defined search space.

The optimization was based on the MSE between the denoised output and the corresponding reference frame. The parameter configuration producing a lower MSE was therefore preferred during the optimization process.

The genetic component of GM-JFSO combined population-based operations such as selection, crossover, and mutation. A mutation rate of 0.1 was used for the implemented genetic operation. The optimization procedure was then configured with the population and iteration parameters, and the genetic component was conducted for multiple generations to explore the parameter space.

3.6. Proposed framework implementation procedure

The implementation procedure followed by the proposed framework is shown below:

Input: Clean endoscopic frames

Y Output: Denoised frames

1. Extract frames from the Hyper-Kvasir video data.

2. Resize each frame to 331 × 331 × 3 and normalize pixel values to [0,1].

3. Generate corresponding noisy frames using the defined degradation procedure.

4. Use NASNetLarge as the feature extraction backbone.

5. Apply the implemented attention mechanism to the selected NASNet feature representation.

6. Reconstruct the denoised RGB frame using the decoder layers.

7. Use GM-JFSO to search the predefined hyperparameter space using MSE as the objective function.

8. Train the proposed model using the selected parameter configuration.

9. Apply the trained model independently to the test frames.

10. Evaluate the denoised frames using PSNR, SSIM, and MSE.

11. Evaluate temporal consistency separately using SSIM between consecutive denoised frames.

3.7. Temporal consistency analysis

The proposed method focuses on spatial denoising and temporal denoising across consecutive denoised frames. To validate the stability, a similarity between adjacent frames metric was introduced to measure the temporal consistency between frames:

$$TC = \frac{1}{N-1}\sum_{t=1}^{N-1} SSIM\left(F_t, F_{t+1}\right) \quad (2)$$

where *Ft* and *Ft + 1* represent the consecutive denoised frames. The higher the temporal consistency values, the smoother the transitions between frames and the reduced flickering artifacts.

Since the proposed network processes individual frames, temporal consistency was evaluated separately after denoising rather than being directly learned through a multi-frame or recurrent architecture. The temporal analysis therefore provides the stability of the denoised output across consecutive frames.

3.8. Comparative evaluation framework

To evaluate the effectiveness of the proposed approach, a comparative analysis was performed using the baseline deep learning models, including NASNet, GoogLeNet, DenseNet, and ResNet. These models provide a comparative evaluation of the proposed framework's denoising performance.

All the models were evaluated under the same experimental conditions using the same datasets and preprocessing steps. The comparison focuses on both spatial quality and temporal consistency, providing an overall assessment of the proposed method's denoising performance.

3.9. Performance metrics

The effectiveness of the proposed framework was evaluated using both spatial and temporal metrics. PSNR and SSIM were used to assess the image quality and structural preservation in the denoised frames. MSE was used to measure the reconstruction error. In addition, the temporal consistency metric was used to evaluate the stability between the frames. This evaluation ensures analysis of the proposed method's performance for both spatial and temporal domains.

4. Results and discussion

4.1. Description of dataset

The proposed framework was evaluated using the publicly available Hyper–Kvasir dataset, a large collection of gastrointestinal endoscopic images and videos developed for medical image analysis research. The dataset contains real-world endoscopic data, which we used to evaluate the proposed framework's performance.

From the Hyper–Kvasir dataset, endoscopic videos were selected and considered for the experimental study, from which a total of 10,000 frames were extracted; 8,000 frames were used for the training stage, and 2,000 frames were used for the testing stage. The extracted video frames were initially converted into individual frames, and these frames were used as input data for the denoising study.

The input videos were then converted into individual frames for preprocessing. Each frame was resized to 331 × 331 pixels to match the requirements of the NASNetLarge-based network. The frames were normalized, and synthetic noise simulation was used to generate the noisy inputs, which is required for supervised denoising. The corresponding clean frames were retained as reference images for performance evaluation of the proposed framework.

During model training, the 8,000 training frames were divided into training and validation subsets using train_test_split with test_size = 0.1 and random_state = 42. Thus, 7,200 frames were used for training, 800 for validation, and 2,000 for testing.

4.2. Implementation details

The proposed framework was implemented using the NASNetLarge-based denoising architecture with a channel attention mechanism and the GM-JFSO optimization technique. The GM-JFSO approach was used to search the predefined parameter space for appropriate model parameters.

The optimization process considers parameters such as learning rate, decoder filter sizes, convolutional filter sizes, and batch size. The main objective function depends on the MSE between the denoised frame and the corresponding reference frame.

After optimization, the model was trained using the Adam optimizer with MSE as the loss function. The model was trained for 100 epochs with a batch size of 8. The baseline deep learning models were trained with the same parameters to maintain consistent experimental settings.

4.3. Quantitative results

The quantitative performance of the proposed model and the selected baseline models was evaluated using PSNR, SSIM, and MSE. Higher PSNR and SSIM indicate better denoising quality and greater structural similarity, whereas lower MSE indicates lower reconstruction error.

The performance of the proposed method was compared with the baseline deep learning models, including NASNet, GoogLeNet, DenseNet, and ResNet, and the results are shown in Figures 2–4. The proposed model achieved higher PSNR and SSIM and lower MSE than the evaluated baseline models.

Figure 2 shows that the proposed model achieved a PSNR of 49.07 dB, compared with 47.40 dB for NASNet, 46.97 dB for GoogLeNet, 45.74 dB for DenseNet, and 45.87 dB for ResNet.

Figure 3 shows the SSIM comparison. The proposed model achieved an SSIM value of 0.938, compared with 0.912 for NASNet, 0.900 for GoogLeNet, 0.860 for DenseNet, and 0.866 for ResNet.

Figure 4 shows the MSE comparison. The proposed model achieved the lowest MSE of 0.584, compared with 1.573 for NASNet, 1.571 for GoogLeNet, 2.081 for DenseNet, and 3.49 for ResNet.

The quantitative results show that the proposed model improves denoising performance across the evaluated metrics, with reduced reconstruction error and improved preservation of structural information in the denoised endoscopic frames.

4.4. Qualitative analysis

The qualitative analysis of the proposed approach was evaluated through visual inspection of denoised frames. The visual analysis is shown in Figure 5. The proposed approach improved noise reduction and better preserved structural details.

4.5. Temporal consistency analysis

Temporal consistency was evaluated by analyzing the consecutive frames in the video sequence. The similarity between the adjacent denoised frames was evaluated using the defined temporal consistency metric as described in Section 3.7. This analysis helps identify flickering artifacts and evaluate the smoothness of transitions between consecutive frames. Temporal consistency across consecutive frames was captured for the provided input video, and the frames are shown in Figure 6. For qualitative illustration of temporal consistency, three consecutive frames (Frames 112, 113, and 114) were selected from the input endoscopic video sequence.

The proposed method was executed on 201 frames, and the temporal SSIM across consecutive frames was 0.8313. Temporal consistency across the frames for the proposed method is shown in graphical form in Figure 7.

The downward spikes were observed in the temporal consistency curve; these variations were due to the abrupt changes between consecutive frames, which were mainly caused by the rapid movement of the endoscope, illumination changes, and the fluids present in the human body. Despite these variations, the proposed method maintained higher overall temporal consistency than the baseline methods, showing improved frame-to-frame stability and reduced flickering artifacts throughout the video sequence.

4.6. Ablation study analysis

An ablation study was conducted to examine the contribution of the channel attention mechanism and GM-JFSO optimization with the proposed framework. For the analysis, a different endoscopic video was selected, and its frames were provided as input after preprocessing. The PSNR and SSIM results obtained from these configurations are shown in Figures 8 and 9, respectively.

The NASNet configuration without channel attention and GM-JFSO achieved a PSNR of 46.54 dB. After integrating the channel attention mechanism with NASNet, the PSNR increased to 47.10 dB. The integration of GM-JFSO with NASNet resulted in a PSNR of 47.62 dB. The configuration combining NASNet, channel attention, and GM-JFSO achieved a PSNR of 48.12 dB.

The corresponding SSIM values for NASNet, NASNet with channel attention, NASNet with GM-JFSO, and the complete proposed framework were 0.891, 0.905, 0.913, and 0.920, respectively.

The ablation results show performance across different model configurations on the randomly selected video. The complete configuration combining NASNet, channel attention, and GM-JFSO achieved the highest PSNR and SSIM among the evaluated configurations, indicating the contribution of the integrated components to the denoising performance.

4.7. Processing time analysis

Figure 10 shows a comparison of processing time between the proposed NASNet model with channel attention and the GM-JFSO framework with baseline deep learning methods. The evaluation was performed using 10, 20, 30, 40, and 50 input frames, and the processing time was measured using the execution time of the respective denoising functions.

The results show that processing time increased with the number of input frames for all evaluated models. This behavior is expected because each additional frame requires a separate denoising operation. The comparison assesses the computational behavior of the proposed framework as input size increases.

The GM-JFSO algorithm was used only during training for hyperparameter optimization; therefore, it did not introduce any additional optimization operations during inference.

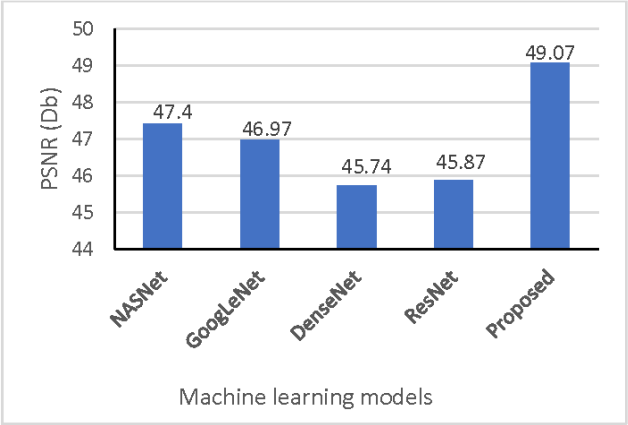


Figure 2. Peak signal-to-noise ratio (PSNR) performance comparison among proposed and baseline models

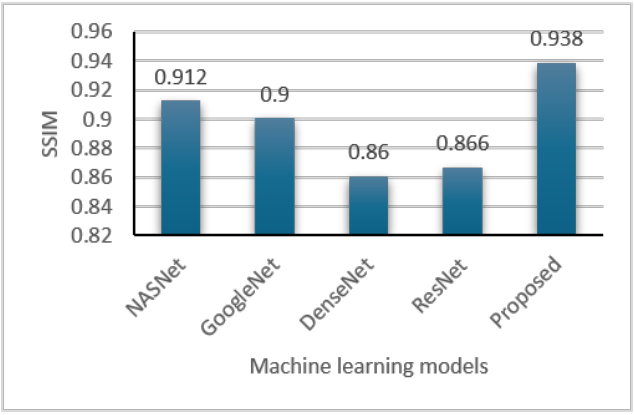


Figure 3. Structural similarity index measure (SSIM) performance comparison among proposed and baseline models

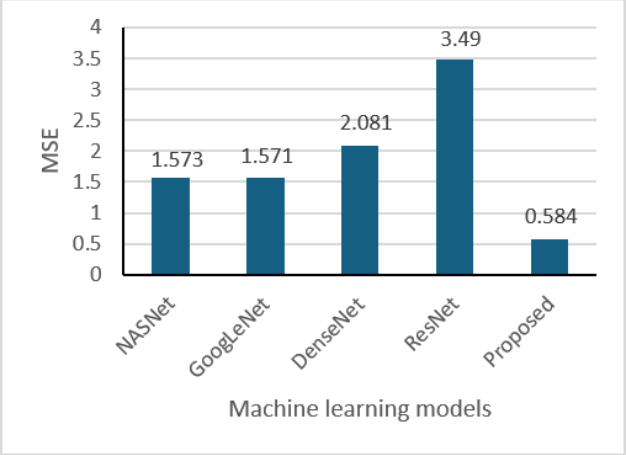


Figure 4. Mean squared error (MSE) performance comparison among proposed and baseline models

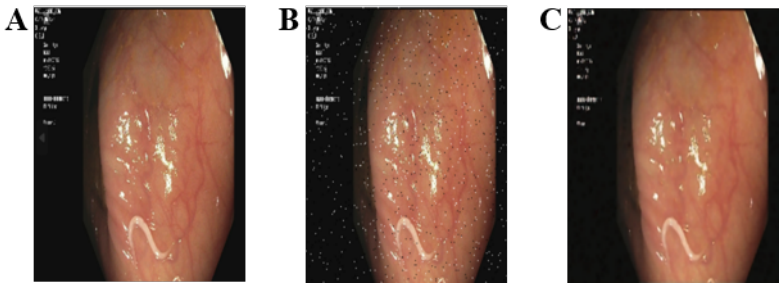


Figure 5. Visual analysis of (a) original, (b) noisy, and (c) denoised frame from the proposed method

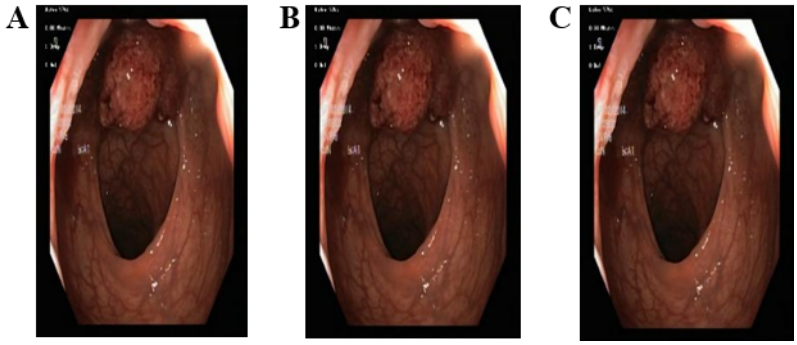


Figure 6. Temporal consistency across frames. (A) Frame 112. (B) Frame 113. (C) Frame 114.

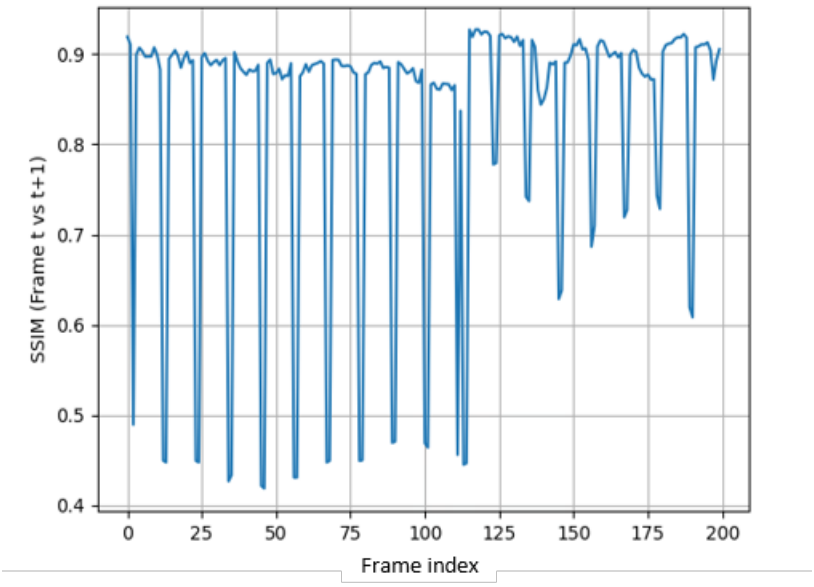


Figure 7. Temporal consistency analysis of the proposed method
Abbreviation: SSIM: Structural similarity index measure.

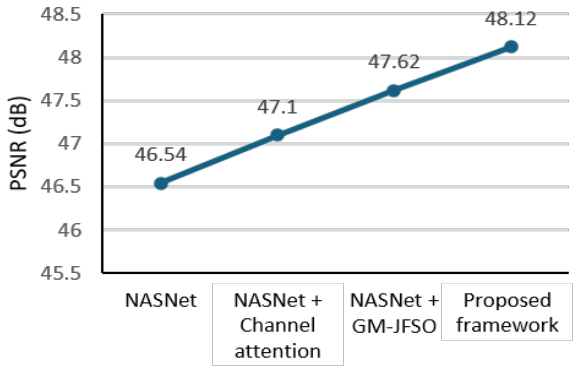


Figure 8. Ablation study analysis in terms of peak signal-to-noise ratio (PSNR)

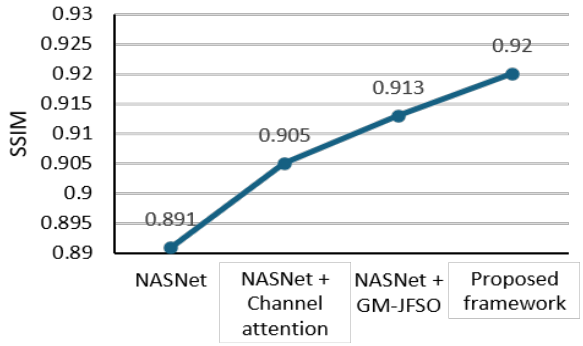


Figure 9. Ablation study analysis in terms of structural similarity index measure (SSIM)

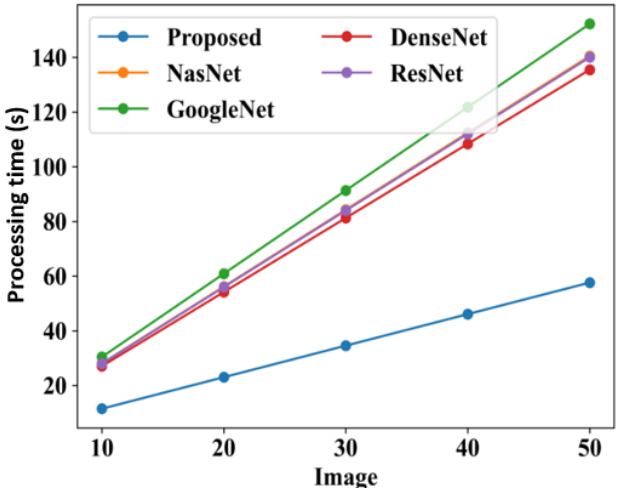


Figure 10. Processing time comparison

4.8. Practical deployment

In the current implementation, the framework was applied offline for post-procedure analysis and enhancement. However, the proposed framework can be extended to support real-time processing using buffered frame sequences, enabling improved denoised visuals with less delay. Real-time deployment may require further model optimization to reduce computational complexity. The processing time analysis shows the computational efficiency of the proposed approach and its potential for practical endoscopic video analysis.

4.9. Limitations

Although the proposed framework has shown superior denoising performance and temporal consistency on the Hyper–Kvasir dataset, several limitations should be acknowledged. The evaluation was conducted using a single publicly available endoscopic dataset. Validation on multi-center datasets and diverse noise conditions would further strengthen the framework's robustness. While the proposed method effectively suppresses noise and preserves structural information, slight smoothing of fine anatomical texture details may occur in highly challenging regions. The current implementation is offline, and optimization is required for real-time clinical deployment.

4.10. Future work

Future work will focus on extending the framework to various endoscopic datasets that contain different noise characteristics and illumination conditions. Further, the work will be conducted to improve texture preservation and the real-time processing capability with advanced optimization strategies. In addition, the proposed framework can be extended to other medical imaging techniques, such as capsule endoscopy, laparoscopic imaging, and ultrasound video analysis.

5. Conclusion

This paper presents an enhanced video denoising framework for endoscopic video frames by integrating NASNet architecture with a channel attention mechanism and a GM-JFSO optimization algorithm. The proposed method focuses on improving the noise reduction while preserving the important structural details in medical video sequences.

In addition to spatial enhancement, the proposed framework evaluates temporal consistency across consecutive frames to ensure smooth transitions and minimize visual artifacts such as flickering. The study also includes a comprehensive evaluation using quantitative metrics and temporal consistency analysis to measure the effectiveness of the proposed approach. The channel attention mechanism focuses on relevant feature representations while suppressing less relevant information, and GM-JFSO optimization enhances the parameter tuning and overall model performance. The comparative evaluation framework provides the advantages of the proposed approach over the baseline deep learning models.

The experimental results highlight the importance of evaluating the denoising performance from multiple perspectives rather than relying on a single metric. The PSNR, SSIM, and MSE results indicate the reconstruction quality of the denoised frames, while the temporal consistency analysis provides an additional measure of frame-to-frame stability. The temporal analysis in the proposed model was performed after frame-wise denoising because the current network processes frames individually rather than multiple consecutive denoised frames. Therefore, the temporal consistency value should be evaluated as a measure of similarity between consecutive denoised frames.

The ablation study analysis evaluated the effect of incorporating the channel attention mechanism and GM-JFSO optimization into the NASNet-based model. The results showed progressive changes in PSNR and SSIM as the individual components were incorporated. The results provide an experimental comparison of the different configurations and support further investigation of the contribution of the integrated components. The findings also indicate that the optimization and feature representation techniques are important considerations when developing the denoising frameworks for endoscopic video data.

The proposed method provides a foundation for further investigation of deep learning-based denoising for medical video applications. Future studies can investigate multi-frame and temporal architectures that directly use information from neighboring frames, along with more diverse degradation models and independently collected clinical video data. Such investigations would help determine the robustness of the proposed framework under variations in motion, illumination, and acquisition conditions.

Acknowledgment

None.

Conflict of interest

The authors declare no competing interests.

Author contributions

Conceptualization: Amulya H. Gopalakrishna, Gopalakrishna Kanathur
Investigation: Amulya H. Gopalakrishna, Gopalakrishna Kanathur
Methodology: Amulya H. Gopalakrishna
Formal analysis: Amulya H. Gopalakrishna, Gopalakrishna Kanathur
Writing – original draft: Amulya H. Gopalakrishna, Gopalakrishna Kanathur
Writing – review & editing: Amulya H. Gopalakrishna, Gopalakrishna Kanathur

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The Hyper–Kvasir dataset is available online at https://datasets.simula.no/hyper-kvasir/.

References

1. Ko WJ, An P, Ko KH, Hahm KB, Hong SP, Cho JY. Image Quality Analysis of Various Gastrointestinal Endoscopes: Why Image Quality Is a Prerequisite for Proper Diagnostic and Therapeutic Endoscopy. *Clin Endosc*. 2015;48(5):374.

   doi: 10.5946/ce.2015.48.5.374

2. Kim GH, Cho YK, Cha JM, Lee SY, Chung IK. Efforts to increase image quality during endoscopy: The role of pronase. *WJGE*. 2016;8(5):267.

   doi: 10.4253/wjge.v8.i5.267

3. Nazir N, Sarwar A, Saini BS. Recent developments in denoising medical images using deep learning: an overview of models, techniques, and challenges. *Micron*. 2024;180:103615.

   doi: 10.1016/j.micron.2024.103615

4. Lin J, Pan Y, Dang Y, Bao Y, Zhuo H. Blood-contaminated endoscopic image restoration based on residual VQ-VAE with cascaded structure. *IET Image Process*. 2022;16(13):3669-3683.

   doi: 10.1049/ipr2.12584

5. Zhang K, Zuo W, Chen Y, Meng D, Zhang L. Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. *IEEE Trans Image Process*. 2017;26(7):3142-3155.

   doi: 10.1109/TIP.2017.2662206

6. Zhang K, Zuo W, Zhang L. FFDNet: toward a fast and flexible solution for CNN-based image denoising. *IEEE Trans Image Process*. 2018;27(9):4608-4622.

   doi: 10.1109/TIP.2018.2839891

7. Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-excitation networks. *IEEE Trans Pattern Anal Mach Intell*. 2020;42(8):2011-2023.

   doi: 10.1109/TPAMI.2019.2913372

8. Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional Block Attention Module. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer; 2018:3-19.

   doi: 10.1007/978-3-030-01234-2_1

9. Tassano M, Delon J, Veit T. DVDNet: A fast network for deep video denoising. In: *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. IEEE; 2019:1805-1809.

   doi: 10.1109/ICIP.2019.8803136

10. Tassano M, Delon J, Veit T. FastDVDNet: towards real-time deep video denoising without flow estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2020:1354-1363.

    doi: 10.1109/CVPR42600.2020.00143

11. Chen H, Jin Y, Xu K, Yang G, Chen H. Multiframe-to-multiframe network for video denoising. *IEEE Trans Multimed*. 2022;23:2507-2519.

    doi: 10.1109/TMM.2021.3077140

12. Maggioni M, Huang Y, Li C, Xiao S, Fu Z, Song F. Efficient multi-stage video denoising with recurrent spatio-temporal fusion. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2021:3466-3475.

    doi: 10.1109/CVPR46437.2021.00347

13. Song M, Zhang Y, Aydın TO. TempFormer: temporally consistent transformer for video denoising. In: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T, eds. *Computer Vision – ECCV 2022*. Springer; 2022:481-496.

    doi: 10.1007/978-3-031-19800-7_28

14. Liang J, Cao J, Fan Y, *et al.* VRT: a video restoration transformer. *IEEE Trans Image Process*. 2024;33:2171-2182.

    doi: 10.1109/TIP.2024.3372454

15. Yue H, Cao C, Liao L, Yang J. RViDeformer: Efficient Raw Video Denoising Transformer With a Larger Benchmark Dataset. *IEEE Trans Circuits Syst Video Technol*. 2025;35(9):8929-8944.

    doi: 10.1109/tcsvt.2025.3553160

16. Ali S, Zhou F, Bailey A, *et al.* A deep learning framework for quality assessment and restoration in video endoscopy. *Med Image Anal*. 2021;68:101900.

    doi: 10.1016/j.media.2020.101900

17. Gómez P, Semmler M, Schützenberger A, Bohr C, Döllinger M. Low-light image enhancement of high-speed endoscopic videos using a convolutional neural network. *Med Biol Eng Comput*. 2019;57(7):1451-1463.

doi: 10.1007/s11517-019-01965-4

18. Nie Z, Xu M, Wang Z, Lu X, Song W. A Review of Application of Deep Learning in Endoscopic Image Processing. *J Imaging*. 2024;10(11):275.

doi: 10.3390/jimaging10110275

19. An Z, Xu C, Qian K, *et al.* EIEN: Endoscopic image enhancement network based on retinex theory. *Sensors*. 2022;22(14):5464.

doi: 10.3390/s22145464

20. Moghtaderi S, Yaghoobian O, Wahid KA, Lukong KE. Endoscopic image enhancement: wavelet transform and guided filter decomposition-based fusion approach. *J Imaging*. 2024;10(1):28.

doi: 10.3390/jimaging10010028

21. Mou E, Wang H, Chen X, *et al.* Retinex theory-based nonlinear luminance enhancement and denoising for low-light endoscopic images. *BMC Med Imaging*. 2024;24:207.

doi: 10.1186/s12880-024-01386-2

22. Borgli H, Thambawita V, Smedsrud PH, *et al.* Hyper-Kvasir: A comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Sci Data*. 2020;7:283.

doi: 10.1038/s41597-020-00622-y

23. Zoph B, Vasudevan V, Shlens J, Le QV. Learning transferable architectures for scalable image recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2018:8697-8710.

doi: 10.1109/CVPR.2018.00907

24. Joy F, Vijayakumar V. A state-of-the-art methodology for real-time surveillance video denoising. In: Smys S, Tavares J, Balas V, Iliyasu A, eds. *Computational Vision and Bio-Inspired Computing. ICCVBIC 2019*. Springer; 2020:985-993.

doi: 10.1007/978-3-030-37218-7_95

25. Ramesh J, Logeshwaran J, Gowri J, Mathew A. The management and reduction of digital noise in video image processing by using transmission-based noise elimination scheme. *ICTACT J Image Video Process*. 2022;13(1):2797-2801.

doi: 10.21917/ijivp.2022.0398

26. Dabov K, Foi A, Katkovnik V, Egiazarian K. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Trans Image Process*. 2007;16(8):2080-2095.

doi: 10.1109/TIP.2007.901238

27. Anwar S, Barnes N. Real image denoising with feature attention. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE; 2019:3155-3164.

doi: 10.1109/ICCV.2019.00325

28. Guo S, Yan Z, Zhang K, Zuo W, Zhang L. Toward convolutional blind denoising of real photographs. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2019:1712-1722.

doi: 10.1109/CVPR.2019.00181

29. Zamir SW, Arora A, Khan S, *et al.* Multi-stage progressive image restoration. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2021:14816-14826.

doi: 10.1109/CVPR46437.2021.01458

30. Zamir SW, Arora A, Khan S, *et al.* Learning enriched features for real image restoration and enhancement. In: Vedaldi A, Bischof H, Brox T, Frahm JM, eds. *Computer Vision – ECCV 2020*. Springer; 2020:492-511.

doi: 10.1007/978-3-030-58595-2_30

31. Liang J, Cao J, Sun GL, Zhang K, Van Gool L, Timofte R. SwinIR: image restoration using swin transformer. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. IEEE; 2021:1833-1844.

doi: 10.1109/ICCVW54120.2021.00210

32. Zamir SW, Arora A, Khan S, Hayat M, Khan FS, Yang MH. Restormer: efficient transformer for high-resolution image restoration. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2022:5718-5727.

doi: 10.1109/CVPR52688.2022.00564

33. Wang ZD, Cun X, Bao JM, Zhou WG, Liu J, Li HQ. Uformer: a general u-shaped transformer for image restoration. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2022:17683-17693.

doi: 10.1109/CVPR52688.2022.01716

34. Jaspers TJM, Boers TGW, Kusters CHJ, *et al.* Robustness evaluation of deep neural networks for endoscopic image analysis: Insights and strategies. *Med Image Anal*. 2024;94:103157.

doi: 10.1016/j.media.2024.103157

35. Chan KCK, Wang X, Yu K, Dong C, Loy CC. BasicVSR: the search for essential components in video super-resolution and beyond. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2021:4947-4956.

doi: 10.1109/CVPR46437.2021.00491

36. Chan KCK, Zhou SC, Xu X, Loy CC. BasicVSR++: improving video super-resolution with enhanced propagation and

alignment. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2022:5972-5981.

doi: 10.1109/CVPR52688.2022.00588

37. Shokeen V, Kumar S, Mathur V, Sharma A, Gupta I, Jain P. ClearNet: auto-encoder based denoising model for endoscopy images. *Indonesian J Elec Eng Comp Sci*. 2025;38(3):1990-2000.

doi: 10.11591/ijeecs.v38.i3.pp1990-2000

38. Jha D, Smedsrud PH, Riegler M, *et al.* Kvasir-SEG: a segmented polyp dataset. In: Ro YM, Cheng W-H, Kim J, *et al.*, eds. *MultiMedia Modeling. MMM 2020*. Springer; 2020:451-462.

doi: 10.1007/978-3-030-37734-2_37

39. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process*. 2004;13(4):600-612.

doi: 10.1109/TIP.2003.819861

40. Rota C, Buzzelli M, Bianco S, Schettini R. Video restoration based on deep learning: a comprehensive survey. *Artif Intell Rev*. 2023;56(6):5317-5364.

doi: 10.1007/s10462-022-10302-5