IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 31, No. 5, May, 2026, pp. 1473-1483
Journal homepage: http://iieta.org/journals/isi

U-Transformer: A Hybrid Convolutional Neural Network–Transformer Framework for Denoising Medical Images Across Imaging Modalities

Rusul Mohammed Neamah*, Noor Kadhim Ayoob, Elaf Ali Abbood, Nada Fadhil Mohammed

Department of Computer Science, College for Science Woman, University of Babylon, Babylon 51002, Iraq

Corresponding Author Email: wsci.rusul.moh@uobabylon.edu.iq

https://doi.org/10.18280/isi.310507

Received: 8 March 2026
Revised: 28 April 2026
Accepted: 8 May 2026
Available online: 31 May 2026

Keywords:
medical image denoising, image restoration, U-Transformer, convolutional neural network–Transformer, multi-modality medical imaging, modality-specific noise simulation, self-attention, perceptual loss

ABSTRACT

Medical image denoising must suppress modality-dependent noise while preserving subtle anatomical and pathological structures. This study presents U-Transformer, a hybrid U-Net and transformer architecture for multi-modality medical image enhancement. The convolutional encoder-decoder extracts multiscale local features through residual blocks and skip connections, whereas a transformer bottleneck models long-range spatial dependencies before residual image reconstruction. A dataset of 21,180 images spanning radiography, magnetic resonance imaging (MRI), computed tomography (CT), ultrasound, dermatology, and microscopy was assembled from public sources. Modality-specific synthetic degradation was introduced using Rician, Gaussian-Poisson, Poisson-Gaussian, speckle, and Gaussian noise models, together with brightness and contrast perturbations. U-Transformer was trained with a composite L1, MS-SSIM, and perceptual loss using an 80:10:10 training, validation, and test split. Performance was assessed using peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and colour difference (ΔE). On radiology, dermatology, and microscopy subsets, the proposed model achieved PSNR values of 42.37, 41.47, and 39.65 dB; SSIM values of 0.9785, 0.9507, and 0.9875; and ΔE values of 2.14, 4.89, and 0.87, respectively. Across the reported comparisons, it obtained the highest PSNR and SSIM and the lowest ΔE for all three image groups. Ablation results further supported the contribution of the transformer bottleneck to restoration performance. These findings indicate that U-Transformer is a promising approach for enhancement of synthetically degraded medical images; validation on raw clinical acquisitions remains necessary.

1. INTRODUCTION

Medical imaging has become one of the indispensable components of the current healthcare, a tool that is necessary in clinical diagnosis, in treatment planning, as well as in post-operative follow-up. The computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, and digital pathology techniques all yield an important percentage of the data on which clinical decisions are made [1]. These images, however, are primarily relied on in terms of their quality in order to be effectively used as a diagnostic tool. Suboptimal image resolution, contrast, or signal-to-noise ratio negatively affects the ability of a clinician to detect fine pathological details and has a direct effect on the quality of the diagnosis and, by extension, patient outcomes [2]. Medical image enhancement can be taken in this context as a discipline of study that has significant scientific and clinical value.

Medical images are nevertheless susceptible to a variety of distortions and effects that limit their quality, despite advances in imaging equipment and acquisition methods. It has been reported that noise due to factors like the reduction of photons in CT or change in temperature due to thermal variation in MRI, is a major impediment to the correct interpretation of anatomical structures especially in low-contrast areas [3]. In the same manner, there can be a lack of contrast between tissues of different densities or magnetic properties, which complicates the definition of lesions and decreases the chances of detecting disease as early as possible. Past experience has shown that such quality issues are not unique to each imaging modality, but are endemic to all phases of medical imaging and, therefore, require strong and universal optimization strategies [4, 5].

The classical image improvement methods have been used to solve the image quality problem in medical imaging over the years. Some of the most widespread algorithms in contrast enhancement of medical images are the histogram equation (HE) and a more adaptive variant, the Contrast-Limited Adaptive Histogram Equation (CLAHE), since they are computationally efficient and can enhance contrast in local areas dynamically [6-8]. In addition, spatial filtering methods that include binary filtering and adaptive denoising schemes including Kalman filters have been shown to be useful in reducing noise, maintaining edge information as shown by new comparative studies [9, 10]. Nevertheless, classical approaches are also typically defined by the fact that they use manually developed criteria and are not adaptable to the heterogeneity of medical images. Multiscale signal analysis Frequency-domain methods, including wave-based denoising

can be complementary to other approaches. These are however also limited to predefined transformation rules and they might not sufficiently reflect the contextual and semantic properties of anatomical structures [11].

The latest developments in deep learning have radically changed the field of medical image enhancement, and complex architectures have proven impressive potentials to enhance the quality of diagnostic images. convolutional neural networks (CNNs) have been shown to be remarkably effective at reducing noise and improving image quality when paired with automated noise-reduction encoders. With a peak signal-to-noise ratio (PSNR) of 35 dB, compared with 30 dB for traditional approaches, their use resulted in a 40% decrease in noise and a 30% increase in resolution [12]. It has been shown that some imaging modalities, especially those based on generative adversarial networks (GANs), can improve cardiac MRI pictures through training. The combined treatment of Gaussian noise, noise reduction, and artifact removal has been the main emphasis of this study [13]. Moreover, contrast has been greatly enhanced through multimodal learning techniques based on periodic competitive generative networks. Better outcomes with low-contrast T2-weighted MRI scans, which rely on features from high-contrast T1-weighted images, serve as evidence of this [14]. Among the best-known high-resolution techniques that show promise for improving the precision and quality of medical images are the SRCNN, VDSR, and ESRGAN algorithms [15]. Additionally, using residual gate mechanisms with comprehensive learning techniques across medical imaging modalities can enhance existing learning strategies by 5-7 dB [16]. Although these learning-based methods have produced significant quality image metrics, their application in the clinical setting should be validated with care through the implementation of various imaging regimes and types of patients to guarantee a sound generalization.

Clinical implications of bettering the quality of medical images go beyond the technical quality improvement; they are intertwined with the endpoints of diagnostic accuracy and healthcare equity. It has been shown empirically that better image quality relates positively to the higher accuracy of diagnostic agreement between radiologists, decreases inter-viewer variability a contributor to the level of doubt experienced in radiology practice [12]. Moreover, computationally efficient optimization approaches, as an effective means of improving the diagnostic capability of the clinic with minimal capital expenditure on equipment upgrades, may be a viable solution to the resource constrained clinical setting [17]. These reflections demonstrate the significance of creating optimization methods, which are both technically sound and can be implemented in clinical practice in practice.

Although there have been critical developments in the conventional as well as deep learning-based image refining techniques, there are still apparent gaps and shortcomings in the published literature. Most modern deep learning methods are tested on single-mode data, and the transferability between modes is also not extensively explored [18, 19]. Moreover, most of the existing improvement frameworks do not have thorough clinical validation research on how the framework affects the subsequent diagnostic task performance. The lack of standardized performance measurement procedures and the partiality in using task-oriented quality metrics are another weakness to make comparisons of studies meaningful. With these constraints, this study will seek to fill these loopholes in the current discourse with a well thought methodology that is empirically based.

Although there has been a notable advance in the field of deep learning-based medical image improving, the majority of the existing approaches use supervised learning with distorted images, which operate on the assumption that the noise distributions are known, like either Gaussian or Poisson noise. Nevertheless, the medical images in the real world may be distorted by complicated and unobservable factors such as uneven lighting, device-dependent distortions, and mixed noise, which reduces the extrapolation of existing models. The main value of this study lies in the development of a new hybrid architecture called U-Transformer, which was specially aimed at solving the difficult issues of improving and denoising medical images. The main contributions of this work may be summarized in the following way: 1. We propose a hybrid architecture that combines CNNs with a lightweight transformer. This approach integrates local spatial features and overall contextual representation. It allows us to preserve long-term structural linkages and restore the microstructure of medicinal tissues. 2. In contrast to standard medical image enhancement programs that add generic Gaussian or Poisson noise, our pipeline uses physically driven, modality-specific noise simulation for six imaging modalities to inject noise patterns that correspond to the real physics of each device.

2. RELATED WORKS

The technologies of medical image enhancement have evolved in an impressive way during the last several decades, developing the traditional signal processing frameworks to the advanced deep learning systems. The evolution is an indication of the technological growth and the increasing clinical need to have quality diagnostic images that will aid in the proper interpretation of different imaging modalities. The evolution path may be divided into two main groups: the traditional approaches, which were prevailing in the sphere until the middle of 2010s, and the modern deep learning solutions, which have been transforming the world of medical image processing since that time.

Histogram-based approaches have always been one of the foundations of medical image enhancement over decades, with CLAHE becoming the most popular method in medical practice. Contrary to the global histogram equalization, CLAHE can work on small blocks of the image, and localized contrast enhancement is applied and a clipping limit mechanism is used to ensure the method does not over-enlarge the blocks [20]. More recent uses have also shown the sustained applicability of CLAHE; Nia and Shih [21] proposed Global-CLAHE (G-CLAHE) a clever version that integrates global and local characteristics in order to balance the demand to retain the overall image context and fine detail in radiography. Their experiments with chest X-rays demonstrated that they performed better than traditional CLAHE especially in the ability to detect subtle anatomical structures without compromising the overall coherence. Besides the concept of histogram, the concept of a hybrid method with a combination of various classical techniques have also received growing interest. Liu and Nguyen [22] showed that synergistic gains can be attained on the basis of integrating CLAHE and wavelet transform decongestion based on nonlocal averages because wavelet analysis can give a multi-resolution analysis, whereas nonlocal averages can

maintain structural information during decongestion. Their performance on the fracture and bone images datasets demonstrated that there was a significant enhancement in diagnostic image quality in comparison to the use of a single method. This paper demonstrates how the traditional approaches, applied thoughtfully can provide the competitive performance of certain clinical tasks. Although they are computationally simple, these techniques are disadvantaged by depending on manual parameters that are not easily changed without manual adjustment, and thus do not allow the methods to be flexible enough to the large clinical heterogeneity.

The advent of deep CNNs has completely redefined the paradigm of medical image enhancement, by allowing learning in a holistic manner of the multifaceted interrelations between degraded images and enhanced images. Zhou et al. [23] have extensively reviewed this change and recorded it, stating that the representational learning emphasis of deep learning as opposed to explicit models has been extremely useful in terms of handling the heterogeneity of medical image distortions. Image enhancement was also one of the areas where deep networks have exhibited a substantial practical effect as the researchers named it among the main applications in conjunction with segmentation and classification tasks. Wang et al. [24] organized a review of the versions of GAN by applying it to medical imaging. They disclosed that conditional generative adversarial networks (cGANs) and CycleGANs were especially useful in enhancing the quality of images that were specific to each imaging mode and images between different imaging modes. They pointed out in their analysis that the twofold benefit of competitive training was that the discriminator trains to offer a learned loss function that more accurately predicts the quality of clinical images than the standard pixel metrics whereas the generator trains to create images under the statistical distribution of high-quality medical scans. Nevertheless, the researchers also noted that there were still persistent problems, such as instability in training, mode collapse, and the inability to create high-resolution images with anatomically meaningful fine details. The achievement of the transformer architectures in natural language processing has led to their application to medical imaging problems, where self-attention mechanism presents benefits in the identification of long-range spatial relationships.

Wang et al. [25] conducted a review of incorporation of transformers into medical image processing pipelines, observing that they are especially useful in applications in which global context sensitivity, inaccessible to only CNN designs, is needed. Swain transformer architecture has been found to be especially useful in image recovery, such as resolution enhancement and noise reduction, and it uses hierarchical feature learning by using offset windows [26].

3. METHODOLOGY

The objective of this study is to create a technique of enhancing medical image quality, with the help of a hybrid deep neural architecture, which is a mixture of U-Net networks and Transformers features. This architecture seeks to solve the recurring problems of medical imaging including noise, but maintain the correct diagnostic information. An improved quality of images is achieved by using a composite loss function to train the model to achieve optical quality and clinical integrity of the model. Figure 1 demonstrates the general block of the proposed method.

The suggested architecture is referred to as U-Transformer improved which use the U-Net architecture and the transformer bottleneck are used because, in contrast to local or channel-only attention, the transformer bottleneck captures long-range dependencies across the whole image. It successfully models patch-wise relations and whole-scene semantics, critical for non-uniform degradations, by balancing global receptive field and computational cost with spatial down sampling. Unlike pure transformers, U-Net offers a multi-scale hierarchical backbone with skip connections that avoid over-smoothing and maintain high-frequency features. This section explains the propose architecture and databases preparation and training:

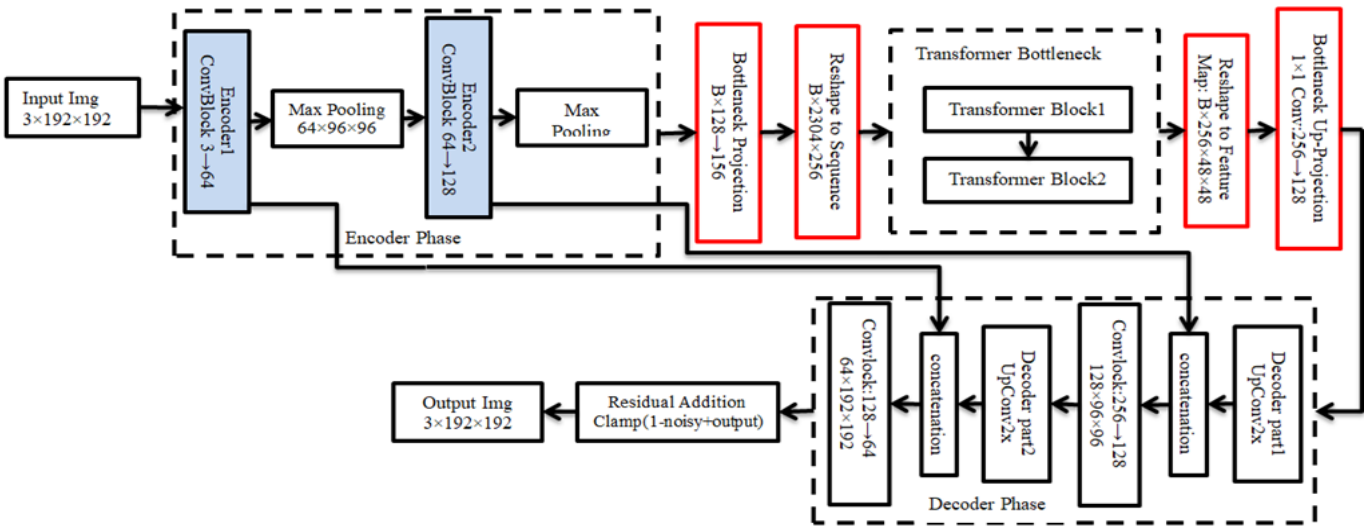

Figure 1. The general architecture of the proposed method

3.1 Encoder phase

The hierarchical multi-stage encoder proposed in the architecture is meant to extract multiscale structural characteristics of medical images. The encoder path is specifically designed to preserve edges and spatial accuracy, in contrast to traditional CNNs, which often result in the loss of fine anatomical details and significantly reduce image size.

The input image is convoluted with two convolutional blocks to obtain the multiscale features. Every block is made up of two sequential layers that are convolutional with Kernel Size (3 × 3), Stride (1), and Padding (1) with a Shortcut path. Both convolutional layers are preceded by BatchNorm and GELU which are used to extract local patterns (vascular edges) as demonstrated in Figure 2. The result of every block is sent to MaxPool2d so as to cut the image by half at every level, enabling the network to view bigger and more global patterns. The features in each level are stored and transmitted directly to the decoder in order to ensure details are fined with Skip Connections.

3.2 Transformer bottleneck phase

The bottleneck section represents the heart of the proposed architecture, where the transition from local convolutional processing to global context modeling takes place. This section is designed to link the spatial features extracted from the encoder via a self-attention mechanism [27], allowing the network to understand long-range interrelationships between medical tissues, as illustrated in Figure 3. To prepare the feature maps for the transformer phase, a 1 × 1 convolutional layer is used to increase the channel depth from 128 to 256 channels. This process, known as bottleneck projection, aims to expand the feature space, giving the transformer greater capacity to represent complex details and nonlinear relationships before initiating the attention process and then pass through two blocks of transformer. Since transformers handle data as sequences, a reshape operation was applied to transform the (48 × 48 × 256) feature maps into a linear sequence with dimensions (2304 × 256). At this stage, 2304 represents the number of spatial pixels (sequence elements), while 256 represents the length of the descriptor vector for each pixel. Positional encoding was then incorporated to ensure the network retained the spatial information and geometric arrangement of the medical organs within the sequence. After processing the data within the transformer blocks using 8-head attention, a reshape to feature map operation was performed to revert the processed sequence to its original spatial form (48 × 48). This step ensured the restoration of the topological order of the features, allowing the decoder to handle them as two-dimensional image maps. In the final stage of the bottleneck, another convolutional layer (1 × 1 Conv) is used to reduce the channel depth from 256 to 128 channels. This up-projection process compresses the focused information extracted by the converter, preparing it for integration with the skip connections coming from the encoder, thus ensuring a balanced flow of information towards the final reconstruction stage.

3.3 Decoder phase

The process of encoder is reversed during the process of decoding and Convolutional Transposed Layers are used to expand the size of feature maps in space. The maps created by the encoder (Connections Skip) are then concatenated with the encoder feature maps that have been upscaled by the decoder. Such combination enables the decoder to draw on both high-level features (the bottleneck) and low-level features (the encoder) hence the ability of retrieving fine image details. The decoder also has each block which corresponds to a convolution block (ConvBlock) like the one used in the encoder. The last layer will be a 1 × 1 convolutional layer which converts the recovered feature maps to the final enhanced image. The step that comes before the generation of the final image is Residual Learning that adds the output with the input by use of Residual Addition. This is an advantage since it shows that the network does not distort the image content but only enhances it. The architecture of the decoding block applied in the model is shown in Figure 4.

Table 1 shows the layers from which the proposed method was built.

Table 1. Layers of the proposed method

Phase	Layers	Filter Size
Encoder	Shortcut Conv	1 × 1
	Encoder Block$_1$: conv$_1$	3 × 3
	MaxPooling	2 × 2
	Encoder Block$_2$: conv$_2$	3 × 3
	MaxPooling	2 × 2
Bottleneck	Projection: Conv	1 × 1
	Transformer Block 8-head Self - Attention	256 Embedding
	Up- Projection: Conv	1 × 1
Decoder	ConvTransposed2d	2 × 2
	Concatenation$_1$ Skip from Encoder$_2$	
	Dcoder Block$_1$: Conv	3 × 3
	ConvTransposed2d	2 × 2
	Concatenation$_2$ Skip from Encoder$_1$	
	Dcoder Block$_2$: Conv	3 × 3
Output	Conv	1 × 1

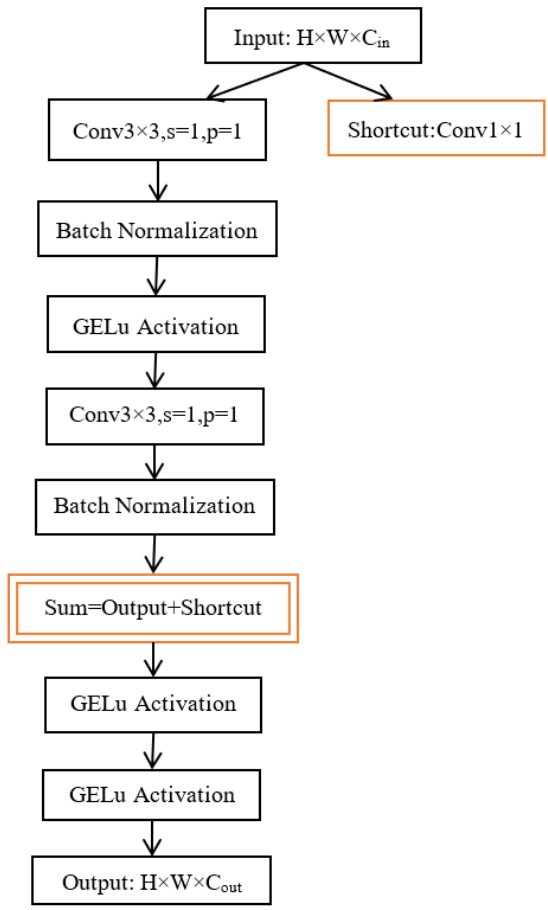

Figure 2. The architecture of the one encoder block

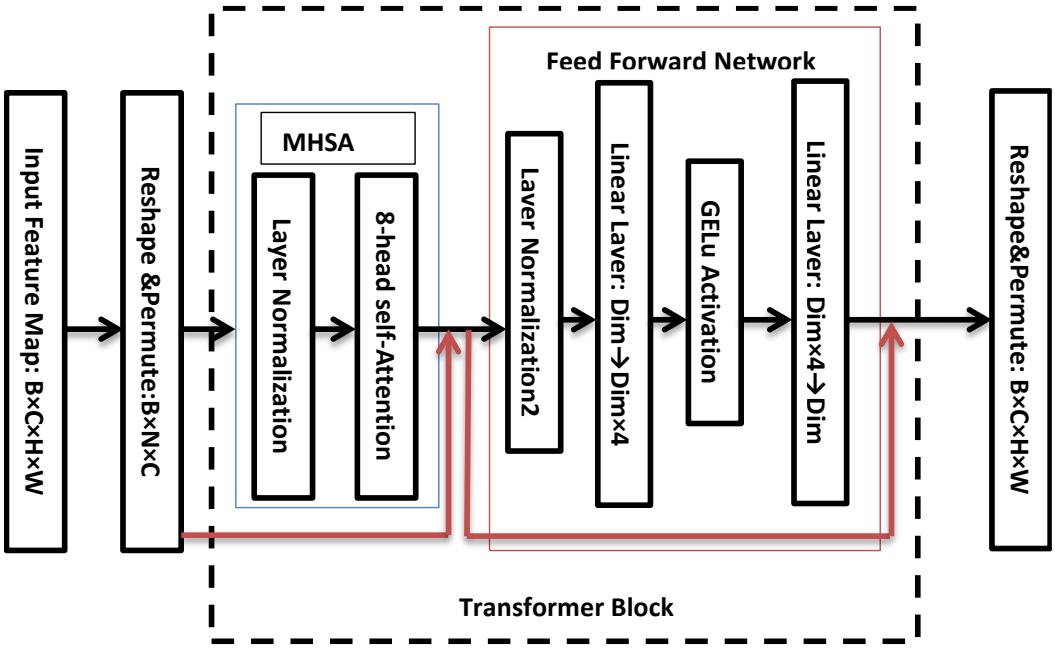

Figure 3. The architecture of the bottleneck with one transformer block

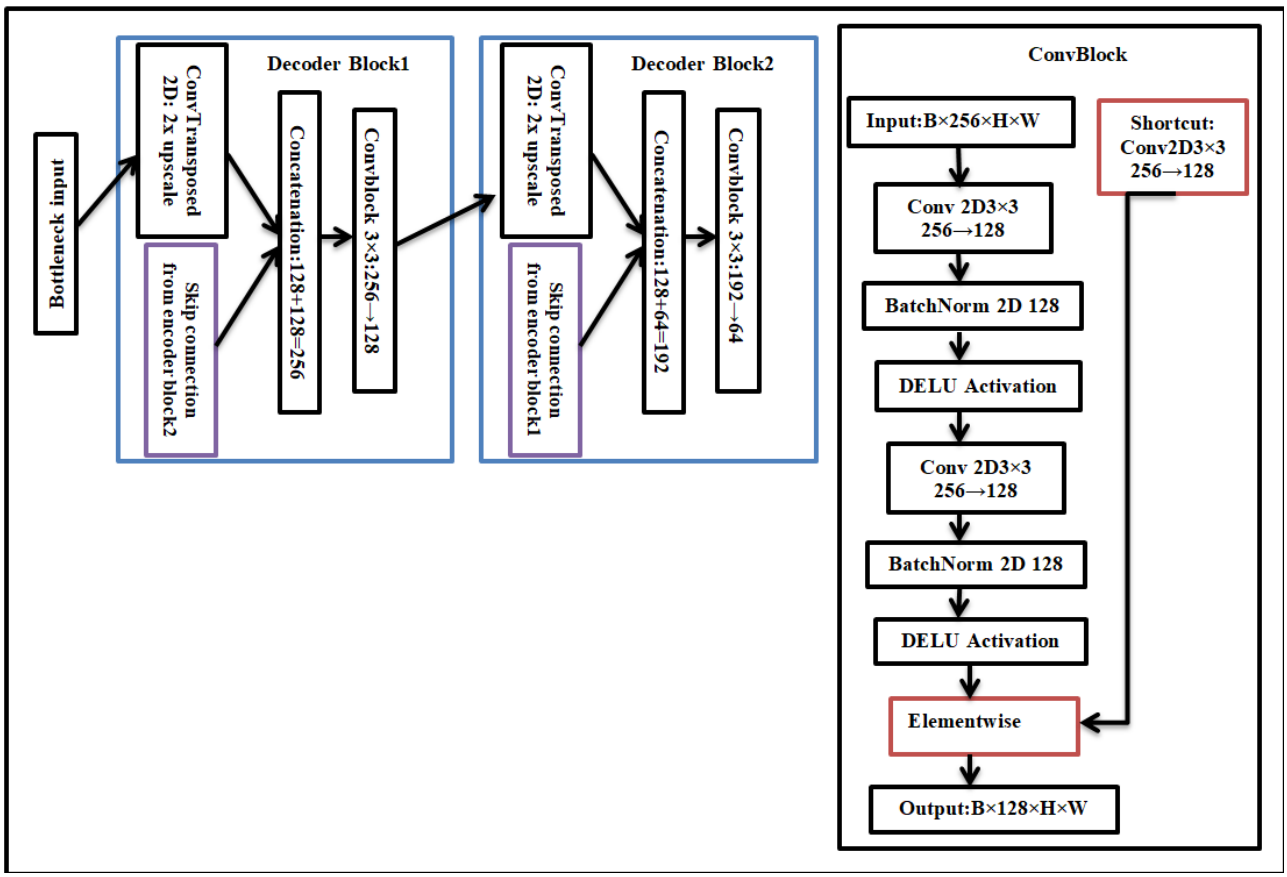

Figure 4. The architecture of the decoder phase

Algorithm 1: Complete U-Transformer Forward Pass
Input: xϵℝ^(B × 3 × 192 × 192) **Output:** yϵℝ^(B × 3 × 192 × 192) **Procedure UTransformer(x, base_dim=64):** **//Encoder** S1←ConvBlock(x, in_channel=3, out_channel= base_dim) e1←MaxPool2d(s1, kernel_size=2, stride=2) s2←ConvBlock(e1,in_channel=base_dim, out_channels=base_dim*2) e2← MaxPool2d(s1, kernel_size=2, stride=2)

```
//Transformer Bottleneck
b, c, h, w←Shap(e2)
e2_proj←Conv2d(e2,kernel_size=1, n_channel=base_dim*2, out_channels=base_dim*4
t_seq←Flatten(e2_proj, start_dim=2)
t_seq←Transpose(t_proj, dim1=1, dim2=2)
for i←1 to 2 do
t_norm←LayerNorm(t_seq)
t_attn,_←MultiheadAttention(t_norm, t_norm, t_norm, embed_dim=256, num_head=8)
t_seq←t_seq+ t_attn
t_norm2←LayerNorm(t_seq)
mlp_hidden←Linear(t_norm2, 256, 1024)
mlp_act←GELU(mlp_hidden)
mlp_out←Linear(mlp_act, 1024,256)
t_seq←t_seq+mlp_out
t_seq←Transpose(t_seq, dim1=1, dim2=2)
t_spatial←View(t_seq, shape=[b, -1, h,w])
t_out←Conv2d(t_spatial, kernel_size=1, in_channel=base_dim*4, out_channels=base_dim*2)
//Decoder
d1_up←ConvTranpose2d(t_out, kernel_size=2, stride=2, in_channel=base_dim*2, out_channels=base_dim*2)
d1_concat←Concatenate([d1_up, s2], dim=1)
d1←ConvBlock(d1_concat, in_channel=base_dim*4, out_channels=base_dim)
d2_up← ConvTranpose2d(d1, kernel_size=2, stride=2, in_channel=base_dim*2, out_channels=base_dim)
d2_concat←Concatenate([d2_up, s1], dim=1)
d2←ConvBlock(d2_concat, in_channel=base_dim*2, out_channels=base_dim)
residual←Conv2d(d2, kernel_size=1, in_channel=base_dim, out_channels=base_dim)
y←CLAMP(x+ residual, min=0.0, max=1.0)
return y
```

4. PREPARATION AND TRAINING OF DATABASES

We were able to collect approximately more than 21,000 medical images from the open web, reflecting a range of technologies used to take them, because there was no specific and easily accessible dataset for the proposed method. Three primary categories were created from the gathered images: i) radiology (X-rays [28], MRI [29], CT [30], ultrasound [31]). ii) dermatology (skin imaging [32]), and iii) microscopy (protein atlas [33], histopathologic [34]) images.

To address the lack of a specific medical images dataset for perceptual improvement research, noise was added to images through a simulation process to create a training dataset. Six modality-specific noise patterns were used, each based on the physical model of the respective medical imaging modality. For example, MRI uses Rician noise with complex distribution is applied at a level of [5, 25] to reflect its complex-valued signal nature in magnetic fields. CT employs a hybrid Gaussian-Poisson model at [10,40], while X-ray radiography applies an inverse Poisson-Gaussian mixture at [8,35]. Ultrasound requires multiplicative Speckle noise at [0.1, 0.3] to simulate multi-path acoustic wave interference. Microscopy and dermatology, due to controlled lighting conditions, utilize Gaussian-Poisson mixtures (1.96%-7.84%) and pure Gaussian noise (1.18%-5.88%), respectively. This noise-centric approach is enhanced by an intensity layer that adjusts brightness and contrast by a factor of [0.9, 1.1]. 21,180 medical images gathered from various imaging modalities make up the entire dataset creation execution. To guarantee complete repeatability, they were divided with an 80-10-10 ratio, producing 16,944 training images, 2,118 images for validation and test respectively.

The optimization criterion applied a well-balanced hybrid loss criterion which is a synergistic combination of pixel resolution, perceptual quality and structural preservation. This synthetic formula dealt with the long-established shortcomings of using error measurements alone per pixel, which do not as a rule represent important perceptual image elements in medical diagnosis [35]. The total loss (Ltotal) was given by:

$$\delta total = \delta L1 + 0.5 \times (1 - \delta MS - SSIM) + 0.01 \times \delta perceptual \quad (1)$$

The L1 component summed the relative distance between the projected and target images and gave useful gradient signals that are not prone to distraction by anomalies as L2 loss. multiscale structural similarity multiscale structural similarity (MS-SSIM) evaluated compliance with a number of measures of resolution - which is paramount to medical imaging in which diagnostically significant features may exist at varying spatial frequencies. The extracted pre-trained VGG19 convolutional layers perceptual term was used to maximize the model outputs according to the learnt semantic representations and thus to improve perceptual realism [36, 37]. The optimization of the parameters used was AdamW variable with discrete weight decay, which was initialized with learning rate of 2×10^{-4} and weight decay of 1×10^{-4} with momentum parameters of 0.9 and 0.999. To avoid rapid convergence in the early training cycles and the optimization of the parameters more accurately in the later training cycles, the learning rate was scheduled based on cosine annealing with Tmax = 100 training cycles used to gradually decrease the learning rate of its initial value to zero [38]. Training was done using 100 cycles with a batch size of 8 and a 80-10-10 split between the training set, validation set and testing set respectively.

All the training process was done in PyTorch 2.0, relying on CUDA-accelerated tensor operations on 16 GB NVIDIA Tesla V100 GPUs. Auto-Accelerated Multi-Precision (AMP) computation utilizing multi-precision training at numerical stability was used [39]. Convergence training required a total

of 22 hours to complete a 100-cycle training schedule and checkpoints were taken every 10 cycles. Based on the standard supervised learning protocols, the model with the minimal loss in verification was kept as the final training network.

5. RESULTS AND DISCUSSION

A disadvantage of this study is that the noisy photos were generated by artificially introducing noise into publicly accessible datasets rather than obtaining raw noisy images directly from medical imaging equipment. While the noise models used in the literature are prevalent in the literature and intended to replicate authentic acquisition settings, they may not comprehensively represent all attributes of real-world noise produced by various imaging equipment, acquisition techniques, and patient circumstances.

Medical images that were gathered were used to assess the effectiveness of the suggested enhancement technique. The study's findings demonstrate an improvement in the quality of the processed medical pictures, as verified by objective assessment metrics. The preservation of exact anatomical features and the lack of aberrations that might skew clinical interpretation are what make these advancements truly significant. Simultaneously, a comparative evaluation was also performed against modern image-enhancement methods, namely the Residual MID [40], a low-light enhancement methodology that uses the dual-illumination-map estimation and incorporating dual-exposure fusion strategies [41], RetinexNet [42], DRAN model that uses dynamic residual attention mechanisms in medical imaging [43], and Deep Perceptual Enhancement [16], which enhances medical image fidelity with the help of perceptual-driven enhancement.

5.1 Quantitative evaluation

A stringent quantitative evaluation is a critical component of determining the objective superiority of image-enhancement methods, which is based on carefully stipulated mathematical value. The effectiveness of the investigated method was measured using three complementary measures, namely the maximum PSNR [44], Structural Similarity Index (SSIM) and the color-difference measure DeltaE [45, 46]. As described in Eqs. (2) and (3) respectively.

$$PSNR = 10\log_{10}(\frac{MAX^2}{MSE}) \tag{2}$$

PSNR is used to measure image quality through the fidelity of pixel-wise image reconstruction by dividing the signal intensity that can be truly attained by the mean-squared error against the original image with respect to which higher PSNR value indicates a better signal preservation and a greater noise reduction [47]. A perceptual statistic called the SSIM measures how much the quality of a processed or compressed image has decreased compared to a perfect reference image. It is predicated on the idea that the human eye is particularly suited to extracting structural details from a picture, such as edges and textures [27]. The chromatic accuracy is measured by the ΔE metric, which is defined as the Euclidean distance in a uniform color space as show in Eq. (3) and ΔE indicates the difference between the reference and deviation, and a smaller ΔE value corresponds to the improvement of a perceptual quality [44].

$$\Delta E^*_{ab} = \sqrt{(L^*_1 - L^*_2)^2 + (a^*_1 - a^*_2)^2 + (b^*_1 - b^*_2)^2} \tag{3}$$

These two measures assess the consistency of signal reconstruction (PSNR) and the truth to perceptual color symbolization (Delta E), as well as provide a comprehensive quantitative description that concurs with the strict mathematical accuracy and the perceptual sensations of the human eye.

As shown in Table 2, the suggested technique consistently surpasses all current methods across all three medical imaging modalities (Radiology, Dermatology, Microscopy) on every criterion.

Radiology: The suggested approach has the lowest DeltaE (2.14), the highest PSNR (42.37), and the highest SSIM (0.9785). The DeltaE improvement is noteworthy when compared to the nearest rival, DRAN (PSNR 41.20, DeltaE 4.23); a decrease from 4.23 to 2.14 indicates an almost half-reduction in the perceived color difference, which is clinically important.

Dermatology: Once more, the suggested approach is superior (PSNR 41.47, SSIM 0.9507, DeltaE 4.89). Even while DeltaE is somewhat greater than in radiology, it is still considerably superior to Deep Perceptual (6.36) and DRAN (9.14). The conversation would be strengthened by a brief mention of the increased variety in skin lesion color.

Microscopic: There has been a significant improvement. Color changes are nearly undetectable to the human eye because of the suggested method's near-perfect SSIM (0.9875) and exceptionally low DeltaE (0.87). This has superb detail retention.

Table 2. Quantitative comparison of the proposed approach with leading optimisation techniques

Method	Radiology Images			Dermatology Images			Microscopic Images		
	PSNR	DeltaE	SSIM	PSNR	DeltaE	SSIM	PSNR	DeltaE	SSIM
Residual MID	36.50	6.34	0.9113	36.18	10.32	0.8851	36.93	7.83	0.8769
DUAL	21.69	6.94	0.8569	16.70	13.26	0.8489	25.21	3.05	0.6670
RetnixNet	20.71	10.89	0.8694	18.28	11.32	0.8568	22.10	8.91	0.8119
DRAN	41.20	4.23	0.9706	40.79	9.14	0.9481	39.36	1.31	0.9735
Deep Perceptual	29.04	3.21	0.8332	23.11	6.36	0.7829	30.69	1.11	0.7976
Proposed	42.37	2.14	0.9785	41.47	4.89	0.9507	39.65	0.87	0.9875

Note: PSNR = peak signal-to-noise ratio; SSIM = Structural Similarity Index.

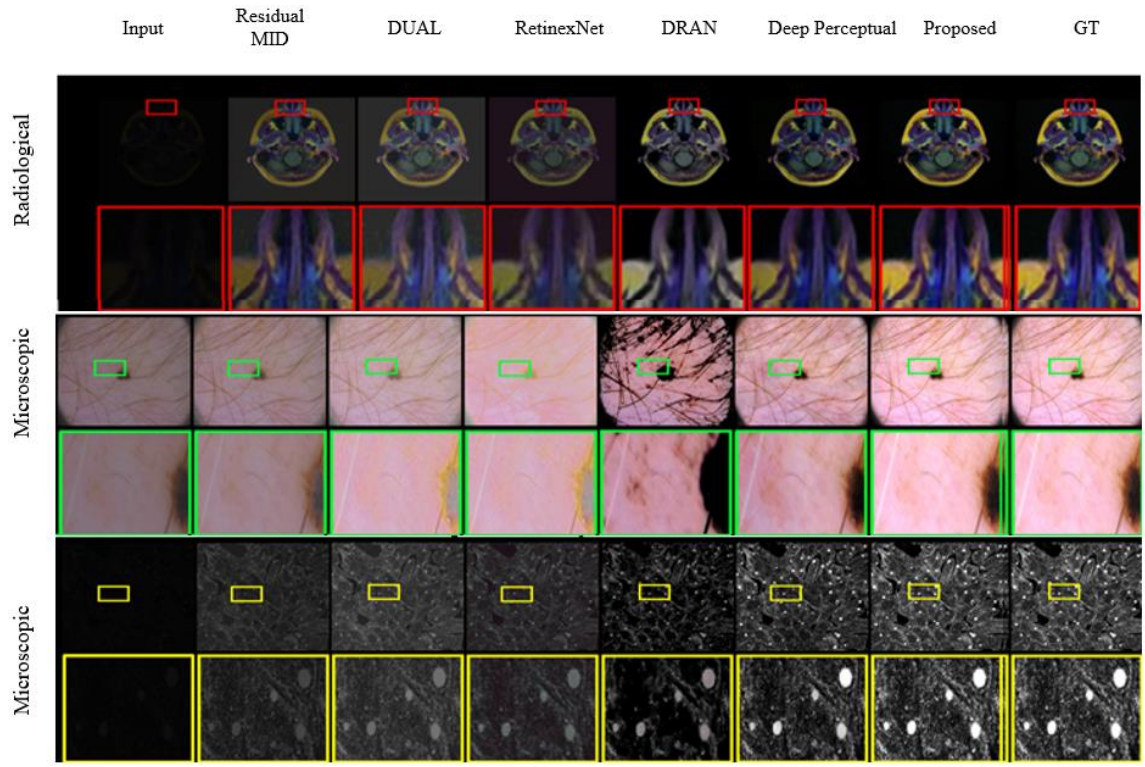

Figure 5. A qualitative image comparative analysis of radiological, microscopic, and microscopic images from top to bottom
Note: From left to right input, Residual MID, DUAL, RetinexNet, DRAN, Deep Perceptual, proposed, and ground truth

5.2 Qualitative evaluation

Whereas quantitative measures provide an objective, mathematically rigorous, measurement, a qualitative visual assessment can never be replaced as far as assessment of perceptual improvement is concerned according to clinical diagnostic criteria. The entire visual comparative analysis of the enhancement results obtained by the proposed U-Transformer are represented in Figure 5 in comparison to five recently existing competing approaches in the format of three medical imaging modalities. The following conclusions can be drawn from a visual comparison of the suggested approach, current approaches, and ground truth (GT) across images. In regions where Residual MID and DUAL lose information (such as around nodules or cell borders), the suggested model accurately recovers anatomical and cellular boundaries. While Deep Perceptual and DRAN exhibit discernible edge blurring, the suggested output in the second microscopic image is almost exactly the same as GT. Subtle grey-level gradations in soft tissue, such as lung parenchyma, are preserved by the suggested technique. RetinexNet, on the other hand, adds artificial contrast and oversaturation, which may distort clinical interpretation. Differentiating between healthy and diseased tissue depends on retaining natural low-contrast information. The suggested method's stain density differs very slightly from GT's in the first microscopic picture. This discrepancy indicates a very modest propensity toward slight contrast enhancement in color-rich regions, but it has little bearing on diagnostic interpretation. In contrast, DRAN clearly shows no saturation in the same area.

6. ABLATION

To isolate the contribution of the transformer module, carried out a number of systematic analytical experiments in the form of a comparison of four network structures (1) a standard U-Net with a standard convolutional encoder-decoder architecture; (2) an U-Net with residual blocks; (3) an U-Net with both residual blocks and skip-attention gates; and (4) propose U-Transformer, which incorporates all the above components. The configurations were trained with the same conditions and were using a composite loss of L1, MS-SSIM, and perceptual terms, and they were optimized using AdamW algorithm with one hundred training epochs on a large corpus of thirty thousand images. Such protocol allowed a fair comparison by standardizing hyper-parameters, and had the same data partitions as demonstrated in Table 3.

Table 3. Ablation analysis of architectural and transformer-based improvements

Network Configuration	Radiology PSNR/DeltaE/SSIM	Dermatology PSNR/DeltaE/SSIM	Microscopy PSNR/DeltaE/SSIM
Baseline U-Net	25.21/ 4.82/ 0.9467	20.43/ 7.29/ 0.8733	27.74/ 1.98/ 0.9452
Residual Blocks	26.98/ 3.17/ 0.9543	21.86/ 6.46/ 0.8739	28.39/ 1.25/ 0.9627
Skip Attention	28.14/ 2.49/ 0.9607	22.58/ 5.78/ 0.8754	30.72/ 1.32/ 0.9689
Proposed	42.37/ 2.14/0.9785	41.47/ 4.89/0.9507	39.65/ 0.87/0.9875

Note: PSNR = peak signal-to-noise ratio; SSIM = Structural Similarity Index.

The empirical findings show that each of the architectural features plays a significant role in the overall performance and the design modifications in the progressively advanced designs support the incremental design logic. Remnant links

always lead to performance enhancement in a variety of modalities due to gradient propagation in hierarchies in the network. Skip-attention connection further improves the performance by selecting to boost salient features and decrease the spread of noise. It is important to note that transformer module incorporation results in strong gains and in radiological applications, where long-range anatomical dependencies need to be modelled in order to enhance the quality of results. Three synergistic mechanisms that include: a global receptive field that allows attention to all spatial locations, adaptive computation that is resource dependent on the complexity of the content, and multiscale implicit feature grouping explain the efficacy of transformers.

7. CONCLUSIONS

The method provides a deep-learning model with a transformer-based extension, which obtains a high level of optimization of multimodal medical images through the simultaneous combination of a global context modeling task and hierarchical feature assessment. The developed U-Transformer showed a lot of gains compared to the current schemes, with an average PSNR gain of 9.2, and a 26.1 decrease in DeltaE over the fields of radiology, microscopy and dermatology. Extensive analysis of exclusion has shown that the integration of transformers has a significant performance contribution (an average PSNR improvement of 1.97 dB, with up to 2.41 dB in radiology) due to three integrated mechanisms: a global receptive domain to model long-range dependencies, adaptive computing to allocate resources depending on content complexity, and implicit multi-scale feature aggregation. The presence of consistent behaviour across the range of different imaging modalities indicates that transformer architectures acquire generalizable optimization principles that are no longer dependent on the noise properties of each particular modality, which is a vital property due to the variety of clinical imaging modalities.

Future research areas are to extend volumetric 3-D imaging with the use of segmented attention mechanisms, investigate self-learning models to overcome a lack of data in dual-training, and apply the framework to a series of diagnostic tasks with multitasking optimization. Nonetheless, our model exhibits distinct limits, including a decline in performance on exceedingly infrequent degradations outside the training distribution, as well as data-gathering issues that limit comprehensive coverage of diversity. To increase the model's resilience, future work will focus on expanding the dataset and addressing these failure scenarios.

REFERENCES

[1] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42: 60-88. https://doi.org/10.1016/j.media.2017.07.005

[2] Rajpurkar, P., Irvin, J., Ball, R.L., Zhu, K., Yang, B., Mehta, H., Duan, T. (2018). Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLOS Medicine, 15(11): e1002686. https://doi.org/10.1371/journal.pmed.1002686

[3] Dong, C., Loy, C.C., He, K., Tang, X. (2015). Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(2): 295-307. https://doi.org/10.1109/TPAMI.2015.2439281

[4] Namozov, A., Cho, Y.I. (2018). An improvement for medical image analysis using data enhancement techniques in deep learning. In 2018 International Conference on Information and Communication Technology Robotics (ICT-ROBOT), Busan, Korea (South), pp. 1-3. https://doi.org/10.1109/ICT-ROBOT.2018.8549917

[5] Goyal, B., Agrawal, S., Sohi, B.S. (2018). Noise issues prevailing in various types of medical images. Biomedical and Pharmacology Journal, 11(3): 1227. https://doi.org/10.13005/bpj/1484

[6] Khudhair, K.T., Najjar, F.H., Waheed, S.R., Al-Jawahry, H.M., Alwan, H.H., Al-khaykan, A. (2023). A novel medical image enhancement technique based on hybrid method. Journal of Physics: Conference Series, 2432(1): 012021. https://doi.org/10.1088/1742-6596/2432/1/012021

[7] Dinh, P.H., Giang, N.L. (2022). A new medical image enhancement algorithm using adaptive parameters. International Journal of Imaging Systems and Technology, 32(6): 2198-2218. https://doi.org/10.1002/ima.22778

[8] Sharma, R., Kamra, A. (2023). A review on CLAHE based enhancement techniques. In 2023 6th International Conference on Contemporary Computing and Informatics (IC3I), Gautam Buddha Nagar, India, pp. 321-325. https://doi.org/10.1109/IC3I59117.2023.10397722

[9] Jayanthi, V., Sivakumar, S. (2022). Image enhancement and de-noising techniques of magnetic resonance images. ASEAN Engineering Journal, 12(3): 137-142. https://doi.org/10.11113/aej.v12.18145

[10] Taassori, M., Vizvári, B. (2024). Enhancing medical image denoising: A hybrid approach incorporating adaptive Kalman filter and non-local means with Latin square optimization. Electronics, 13(13): 2640. https://doi.org/10.3390/electronics13132640

[11] Okuwobi, I.P., Ding, Z., Wan, J., Jiang, J. (2023). SWM-DE: Statistical wavelet model for joint denoising and enhancement for multimodal medical images. Medicine in Novel Technology and Devices, 18: 100234. https://doi.org/10.1016/j.medntd.2023.100234

[12] Sundarrajan, M., Choudhry, M.D., Biju, J., Krishnakumar, S., Rajeshkumar, K. (2024). Enhancing low-light medical imaging through deep learning-based noise reduction techniques. Indian Journal of Science and Technology, 17(34): 3567-3579. https://doi.org/10.17485/IJST/v17i34.2489

[13] Jiang, Y., Cui, L., Jiang, B., Zhao, X., Chai, S. (2024). Cardiac MRI image enhancement based on GAN network. In 2024 43rd Chinese Control Conference (CCC), Kunming, China, pp. 8309-8315. https://doi.org/10.23919/CCC63176.2024.10661588

[14] Naseem, R., Islam, A.J., Cheikh, F.A., Beghdadi, A. (2022). Contrast enhancement: Cross-modal learning approach for medical images. Electronic Imaging, 34: 1-6. https://doi.org/10.2352/EI.2022.34.10.IPAS-344

[15] Yamashita, K., Markov, K. (2020). Medical image

enhancement using super resolution methods. In International Conference on Computational Science, 12141: 496-508. https://doi.org/10.1007/978-3-030-50426-7_37

[16] Sharif, S.M.A., Naqvi, R.A., Biswas, M., Loh, W.K. (2022). Deep perceptual enhancement for medical image analysis. IEEE Journal of Biomedical and Health Informatics, 26(10): 4826-4836. https://doi.org/10.1109/JBHI.2022.3168604

[17] Slonopas, A., Beatty, A., Djajalaksana, Y. (2024). Applying reservoir computing and machine learning techniques for image enhancement in biomedical imaging. In 2024 International Conference on Smart Applications, Communications and Networking (SmartNets), Harrisonburg, VA, USA, pp. 1-7. https://doi.org/10.1109/SmartNets61466.2024.10577705

[18] Li, Y., Sixou, B., Peyrin, F. (2021). A review of the deep learning methods for medical images super resolution problems. IRBM, 42(2): 120-133. https://doi.org/10.1016/j.irbm.2020.08.004

[19] Singh, N.T., Kaur, C., Chaudhary, A., Goyal, S. (2023). Preprocessing of medical images using deep learning: A comprehensive review. In 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), Trichy, India, pp. 521-527. https://doi.org/10.1109/ICAISS58487.2023.10250462

[20] Vidhya, G.R., Ramesh, H. (2017). Effectiveness of contrast limited adaptive histogram equalization technique on multispectral satellite imagery. In Proceedings of the International Conference on Video and Image Processing, New York, NY, pp. 234-239. https://doi.org/10.1145/3177404.3177409

[21] Nia, S.N., Shih, F.Y. (2024). Medical X-ray image enhancement using global contrast-limited adaptive histogram equalization. International Journal of Pattern Recognition and Artificial Intelligence, 38(12): 2457010. https://doi.org/10.1142/S0218001424570106

[22] Liu, X., Nguyen, T.D. (2024). Medical images enhancement by integrating CLAHE with wavelet transform and non-local means denoising. Academic Journal of Computer and Information Science, 7: 52-58. https://doi.org/10.25236/AJCIS.2024.070108

[23] Zhou, S.K., Greenspan, H., Davatzikos, C., Duncan, J.S., Van Ginneken, B., Madabhushi, A., Summers, R.M. (2021). A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE, 109(5): 820-838. https://doi.org/10.1109/JPROC.2021.3054390

[24] Wang, S., Zhou, X., Li, C., Wang, S., Li, Y., Tan, T., Zheng, H. (2025). Generative artificial intelligence in medical imaging: Foundations, progress, and clinical translation. Research, 8: 1029. https://doi.org/10.34133/research.1029

[25] Wang, J., Zhu, H., Wang, S.H., Zhang, Y.D. (2021). A review of deep learning on medical image analysis. Mobile Networks and Applications, 26(1): 351-380. https://doi.org/10.1007/s11036-020-01672-7

[26] Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R. (2021). Swinir: Image restoration using swin transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, pp. 1833-1844. https://doi.org/10.1109/ICCVW54120.2021.00210

[27] Neamah, R.M., Al-Asadi, T.A. (2025). Transformer and spatial convolution-based visual and infrared image fusion. AIP Conference Proceedings, 3264(1): 030002. https://doi.org/10.36478/ajit.2016.2756.2762

[28] Rajpurkar, P. (2017). CheXpert: A large chest x-ray dataset and competition. Stanford ML Group. https://stanfordmlgroup.github.io/competitions/chexpert.

[29] Buda, M., Saha, A., Mazurowski, M.A. (2019). Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. Computers in Biology and Medicine, 109: 218-225. https://doi.org/10.1016/j.compbiomed.2019.05.002

[30] Yang, X., He, X., Zhao, J., Zhang, Y., Zhang, S., Xie, P. (2020). COVID-CT-dataset: A CT scan dataset about COVID-19. arXiv Preprint arXiv:2003.13865. https://doi.org/10.48550/arXiv.2003.13865

[31] Baby, M., Jereesh, A.S. (2017). Automatic nerve segmentation of ultrasound images. In 2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, pp. 107-112. https://doi.org/10.1109/ICECA.2017.8203654

[32] Rezvantalab, A., Safigholi, H., Karimijeshni, S. (2018). Dermatologist level dermoscopy skin cancer classification using different deep learning convolutional neural networks algorithms. arXiv Preprint arXiv:1810.10348. https://doi.org/10.48550/arXiv.1810.10348

[33] Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Ponten, F. (2010). Towards a knowledge-based human protein atlas. Nature Biotechnology, 28(12): 1248-1250. https://doi.org/10.1038/nbt1210-1248

[34] Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M. (2018). Rotation equivariant CNNs for digital pathology. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 210-218. https://doi.org/10.1007/978-3-030-00934-2_24

[35] Wang, Z., Liu, M., Cheng, X., Zhu, J., Wang, X., Gong, H., Xu, L. (2023). Self-adaption and texture generation: A hybrid loss function for low-dose CT denoising. Journal of Applied Clinical Medical Physics, 24(9): e14113. https://doi.org/10.1002/acm2.14113

[36] Johnson, J., Alahi, A., Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In European Conference on Computer Vision, Springer, Cham, 9909: 694-711. https://doi.org/10.1007/978-3-319-46475-6_43

[37] Tatlıcan, D., Apaydin, N.N., Yaman, O., Karakose, M. (2025). Crowd density estimation via a VGG-16-based CSRNet model. Information Dynamics and Applications, 4(2): 66-75. https://doi.org/10.56578/ida040201

[38] Loshchilov, I., Hutter, F. (2017). Decoupled weight decay regularization. arXiv Preprint arXiv:1711.05101. https://doi.org/10.48550/arXiv.1711.05101

[39] Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Wu, H. (2017). Mixed precision training. arXiv Preprint arXiv:1710.03740. https://doi.org/10.48550/arXiv.1710.03740

[40] Jifara, W., Jiang, F., Rho, S., Cheng, M., Liu, S. (2019). Medical image denoising using convolutional neural network: A residual learning approach. The Journal of

Supercomputing, 75(2): 704-718. https://doi.org/10.1007/s11227-017-2080-0
[41] Zhang, Q., Nie, Y., Zheng, W.S. (2019). Dual illumination estimation for robust exposure correction. Computer Graphics Forum, 38(7): 243-252. https://doi.org/10.1111/cgf.13833
[42] Wei, C., Wang, W., Yang, W., Liu, J. (2018). Deep retinex decomposition for low-light enhancement. arXiv Preprint arXiv:1808.04560. https://doi.org/10.48550/arXiv.1808.04560
[43] Sharif, S.M.A., Naqvi, R.A., Biswas, M. (2020). Learning medical image denoising with deep dynamic residual attention network. Mathematics, 8(12): 2192. https://doi.org/10.3390/math8122192
[44] Huynh-Thu, Q., Ghanbari, M. (2008). Scope of validity of PSNR in image/video quality assessment. Electronics Letters, 44(13): 800-801. https://doi.org/10.1049/el:20080522
[45] Luo, M.R., Cui, G., Rigg, B. (2001). The development of the CIE 2000 colour-difference formula: CIEDE2000. Color Research and Application, 26(5): 340-350. https://doi.org/10.1002/col.1049
[46] Sharma, G., Wu, W., Dalal, E.N. (2005). The CIEDE2000 colour-difference formula: Implementation notes, supplementary test data, and mathematical observations. Color Research and Application, 30(1): 21-30. https://doi.org/10.1002/col.20070
[47] Hore, A., Ziou, D. (2010). Image quality metrics: PSNR vs. SSIM. In 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, pp. 2366-2369. https://doi.org/10.1109/ICPR.2010.579