IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 29 April 2024, accepted 14 May 2024, date of publication 20 May 2024, date of current version 20 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3402953

RESEARCH ARTICLE

Saliency-Aware Deep Learning Approach for Enhanced Endoscopic Image Super-Resolution

MANSOOR HAYAT[1], (Graduate Student Member, IEEE), AND SUPAVADEE ARAMVITH[2], (Senior Member, IEEE)

[1]Department of Electrical Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok 10330, Thailand
[2]Multimedia Data Analytics and Processing Research Unit, Department of Electrical Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok 10330, Thailand

Corresponding author: Supavadee Aramvith (supavadee.a@chula.ac.th)

This work was supported by the Graduate Scholarship Programme for ASEAN or Non–ASEAN Countries and Ratchadapiseksompotch Fund Chulalongkorn University.

ABSTRACT The adoption of Stereo Imaging technology within endoscopic procedures represents a transformative advancement in medical imaging, providing surgeons with depth perception and detailed views of internal anatomy for enhanced diagnostic accuracy and surgical precision. However, the practical application of stereo imaging in endoscopy faces challenges, including the generation of low-resolution and blurred images, which can hinder the effectiveness of medical diagnoses and interventions. Our research introduces an endoscopic image SR model in response to these specific. This model features an innovative feature extraction module and an advanced cross-view feature interaction module tailored for the intricacies of endoscopic imagery. Initially trained on the SCARED dataset, our model was rigorously tested across four additional publicly available endoscopic image datasets at scales 2, 4, and 8, demonstrating unparalleled performance improvements in endoscopic SR. Our results are compelling. They show that our model not only substantially enhances the quality of endoscopic images but also consistently surpasses other existing methods like E-SEVSR, DCSSRNet, and CCSBESR in all tested datasets, in quantitative measures such as PSNR and SSIM, and in qualitative evaluations. The successful application of our SR model in endoscopic imaging has the potential to revolutionize medical diagnostics and surgery, significantly increasing the precision and effectiveness of endoscopic procedures. The code will be released on GitHub and can be accessed at https://github.com/cu-vtrg-lab/Saliency-Aware-Deep-Learning-Approach-for-Enhanced-Endoscopic-Image-SR.

INDEX TERMS Robotic surgery, stereo endoscopic surgical imaging, SR, surgical instruments.

I. INTRODUCTION

The progression of digital imaging technology has markedly enhanced visual quality in various sectors. The advent of high-resolution imaging, from black-and-white to 8K resolutions, underscores the pivotal role of pixel density in defining image clarity. Hussain et al. [51] present an innovative approach utilizing a hybrid deep learning model Combining AlexNet and SVM to improve the classification of fetal health status from cardiotocographic data will significantly enhance

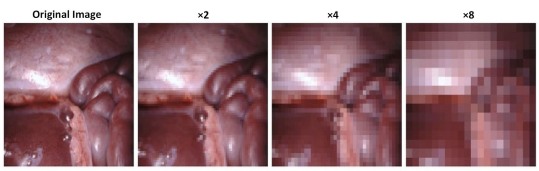

FIGURE 1. Visual comparison of endoscopic image quality at various downsampled resolutions: Original, ×2, ×4, and ×8.

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wei.

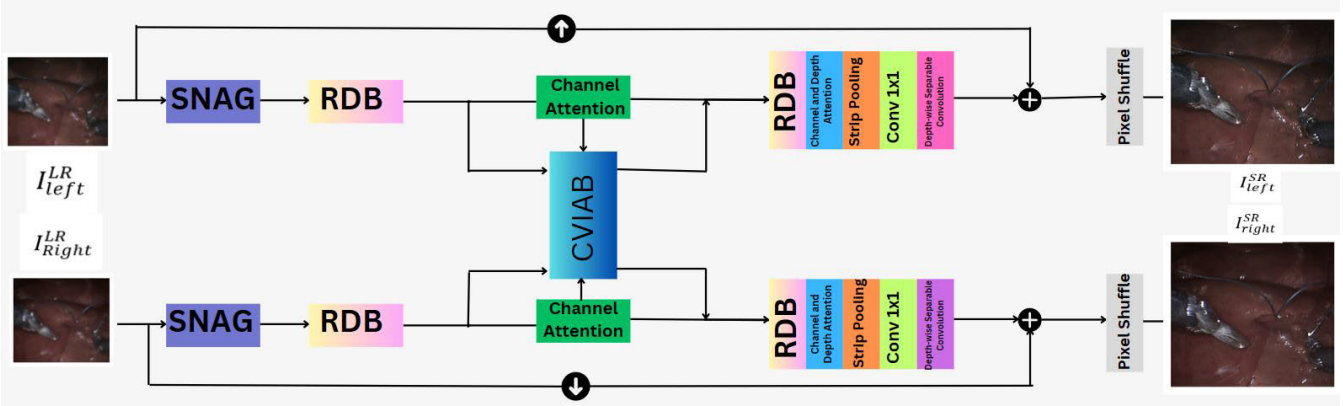

FIGURE 2. The proposed network architecture is illustrated with I^{LR} symbolizing the input low-resolution images on the left side, and I^{SR} denoting the output reconstructed image on the right side. This network is designed to process low-resolution endoscopic images and produce high-quality, super-resolved images as output.

real-time clinical decision-making. Raza et al. [52] developed a hybrid deep learning model, DeepTumorNet, leveraging convolutional neural network architectures to enhance the accuracy of brain tumor classification from MRI images. In their extensive evaluations, they demonstrated superior performance over traditional models. However, the need for better image quality faces significant hurdles, where blurriness, noise, and the loss of critical details compromise the integrity of visual data. Despite offering depth and enriched detail through its dual-viewpoint approach, stereo imaging is notably susceptible to these challenges, owing to its intricate spatial and temporal dynamics. This predicament has ignited interest in super-resolution (SR) techniques to reconstruct high-quality visuals from lower-resolution sources. Jiang et al. introduce the Adaptive-Threshold-Based Multi-Model Fusion Network (ATMFN) [45], an innovative approach that employs CNN-, GAN-, and RNN-based super-resolvers in an ensemble learning framework to enhance the quality of compressed face hallucination images by leveraging the advantages of diverse learning models. Jiang et al. [46] present the Dual-Path Deep Fusion Network (DPDFN). This robust system integrates global and local facial features through a novel dual-path architecture, significantly enhancing the SR of face images without the need for extensive facial prior information. In the EDiffSR framework, Xiao et al. [47] develop an efficient diffusion probabilistic model tailored for remote sensing image SR, achieving superior performance by integrating a conditional prior enhancement module that leverages informative cues from low-resolution images.

In the domain of endoscopic surgery—a staple for minimally invasive procedures—the adoption of stereo cameras has been pivotal in transcending the depth perception and field of view limitations presented by traditional single-camera setups [1], [2]. Stereo endoscopic imagery, with its dual-viewpoint depth cues, significantly outpaces the single-camera modality in delivering superior visual information [3]. Nonetheless, the difficulties of the surgical milieu, such as constrained operational spaces and fluctuating lighting conditions, invariably impair the quality and resolution of stereo endoscopic visuals, thus impeding essential analytical tasks, including image classification, segmentation, and reconstruction [3], [4]. As illustrated in Fig 1, the progressive degradation of image quality from left to right significantly hinders the ability of doctors and medical practitioners to accurately interpret the visual information, which is crucial for effective diagnosis and treatment.

In addressing these intricacies, our study pioneers a novel stereo endoscopic image SR paradigm. Our contributions are deliberate and multifaceted:

- We debut an advanced feature extraction module specifically engineered to distill fine-grained features from stereo endoscopic images, elevating the SR reconstruction's accuracy.
- A novel cross-view feature interaction module is introduced. It refines the processing of stereo scene images by capitalizing on the depth and spatial disparities across viewpoints to augment detail visibility and depth discernment.
- Our investigation extends to SR at scales 2, 4, and 8, marking a pioneering stride in SR research to evaluate and elucidate the impact at scale 8. Our research distinguishes itself by extending SR investigation to unprecedented scales—particularly scale 8—making this the first study within the field to assess and establish the transformative impact of such a high scaling factor in endoscopic image SR.
- The robustness and adaptability of our approach are underscored by comprehensive testing across five publicly available endoscopic image datasets, setting a new precedent in the literature for such extensive empirical validation.

By tailoring our efforts to the specific demands of stereo endoscopic image SR, we aim to fortify the visual quality of endoscopic procedures and lay the groundwork for subsequent innovations in medical imaging technology.

II. LITERATURE REVIEW

This section reviews SR techniques pertinent to our study: single-image SR [13], stereo-image SR [14], video SR [15], and stereo endoscopic SR [5], with a focus on enhancing endoscopic imagery. Single-image SR (SISR) aims to improve the resolution of individual images, often leading to suboptimal results for dynamic scenes such as those found in endoscopy. Multiple Image SR (MISR) addresses this by using various low-resolution (LR) images to generate a high-resolution (HR) image, thereby improving quality.

These methods typically involve analyzing LR and HR image pairs to learn a transformation to HR. Depending on the training data set, the techniques vary, catering to general or specific types of images, such as medical images. Sparse coding is a notable method for example-based SR.

Recent advances have utilized deep convolutional neural networks for SR reconstruction. Hu et al. [16] refined the network architecture to enhance performance and simplify training. Other methods have focused on improving feature relationships and overall image quality by incorporating context into the network. Despite progress, ongoing research is important for further improvements in SR techniques.

A. SINGLE IMAGE SR

Single Image SR (SISR) has witnessed substantial progress with the integration of deep learning techniques, notably in enhancing reconstruction accuracy [17], [18], [19]. The SENext model employs squeeze-and-excitation blocks and various skip connections to decrease computational demands while improving feature processing [20]. This approach, together with the Very Deep SR Network (VDSR) by Kim et al. [21], the Residual Dense Network (RDN) by Zhang et al. [22], and architectures like RCAN [23], RNAN [24], and SAN [25], illustrate the growing complexity of SR networks. Muhammad et al. introduced an innovative architecture that reduces network parameters while boosting speed and image quality, signifying a pivotal advancement in SISR [26]. Recent methodologies have further expanded the scope of SR techniques. The ATMFN model [45] combines diverse deep learning models through adaptive threshold-based fusion, significantly improving face hallucination by leveraging their collective strengths. The TTST model [48] innovates within transformer architectures by incorporating a top-k token selective mechanism, optimizing the attention process in SR tasks and reducing computational overhead.

B. STEREO IMAGE SR

Stereo image SR is critical for extracting high-quality images from stereo pairs, leveraging the rich stereo information. Bhavsar and Rajagopalan [14] and others have explored depth and image enhancement methods, employing iterative processes. CNN-based techniques have marked a significant evolution in this field, with Jeon et al. [8] addressing parallax challenges and Wang et al. [27] focusing on cross-view information capture through their Parallax-Attention mechanism.

Yan et al. [7] have furthered this by adapting domain-specific networks to utilize cross view data more effectively, while Xu et al. [28] and Chu et al. [29] have refined CNN methods, focusing on accurate disparity maps and information integration. These methodologies are central to advancing stereo SR, demonstrating the potential for nuanced disparity and detail capture to significantly enhance the SR process and offer a comprehensive solution for high-fidelity SR imaging in stereoscopic applications.

C. VIDEO SR

VSR distinguishes itself from SISR by using temporal correlations between frames for reconstruction [23], [30], [31], [32], with methods often involving frame alignment through motion compensation. Despite challenges in optical flow estimation [33], [34], approaches like that of Wang et al. [12] combine alignment with attention modules for improved results. Additionally, the Local-Global Temporal Difference learning network [49] presents a novel approach by utilizing both short-term and long-term temporal differences for effective satellite video SR.

D. STEREO ENDOSCOPIC SR

Stereo Video SR has been explored to address challenges like deblurring and adapting stereo content to various screen sizes, utilizing 3D scene flow and depth information for enhanced performance [35], [36], [37]. In endoscopic surgery, SR techniques have evolved to improve visual clarity, from the Minimally Invasive Surgery SR in 2011 [41] to recent developments focusing on attention mechanisms for detail enhancement and image reconstruction [5], [40]. Hayat et al. [44] proposed a novel algorithm for endoscopic image SR and surgical instrument segmentation. This progression reflects a commitment to advancing image quality in endoscopy, balancing real-time applicability, environmental adaptability, and computational demands.

III. PROPOSED METHODOLOGY

Detailed Description of the Network Architecture: Our network architecture is designed to address the complex requirements of enhancing stereo endoscopic images through SR techniques. The architecture comprises several critical components: the StereoNet Attention Gate (SNAG), Residual Dense Block (RDB), Cross-View Interactive Attention Block (CVIAB), Reconstruction Block, and Pixel Shuffle. Each element plays a vital role in processing and improving the image data efficiently.

The model $\mathcal{J}$, parameterized by ϕ, generates super-resolved (SR) images $I_k^{(\text{left,SR})}$ and $I_k^{(\text{right,SR})}$ from low-resolution (LR) frames for both the left $\left(I_{(k-m)}^{(\text{left,LR})}, \ldots, I_{(k+m)}^{(\text{left,LR})}\right)$ and right $\left(I_{(k-m)}^{(\text{right,LR})}, \ldots, I_{(k+m)}^{(\text{right,LR})}\right)$ views. This approach allows for an enhanced focus on areas of the image that hold the most diagnostic value, which is essential for

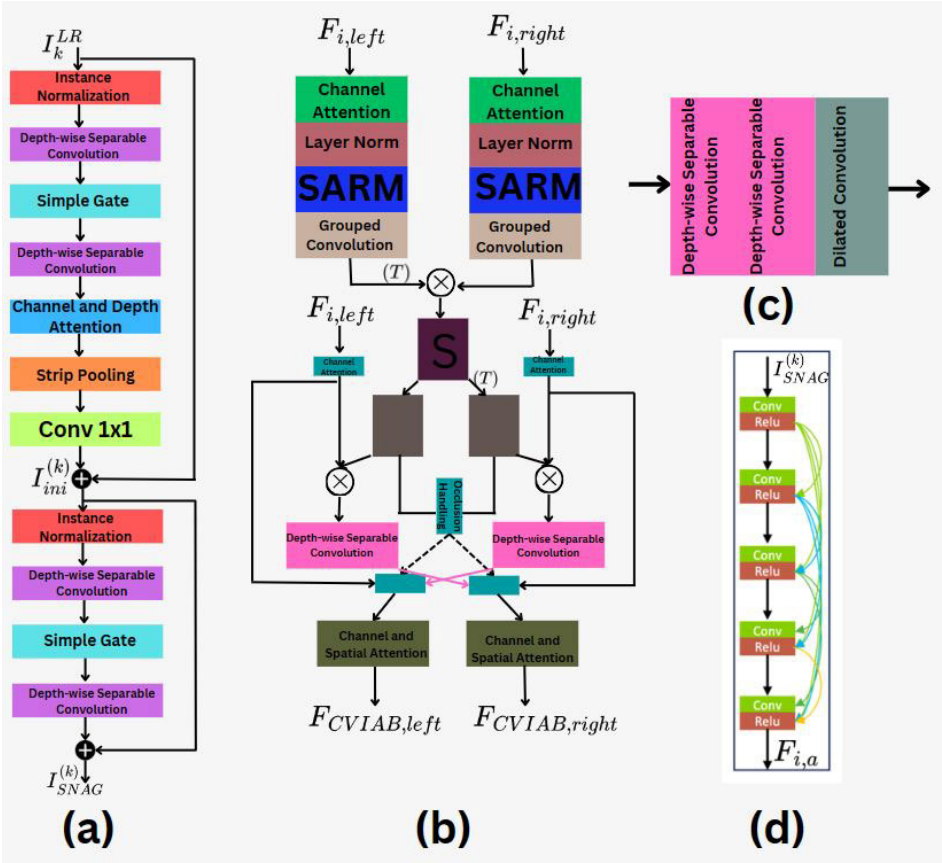

FIGURE 3. (a): StereoNet Attention Gate (SNAG), (b):Cross-View Interactive Attention Block (CVIAB), (c): Stereo Attentional Residual Module (SARM), (d): Residual Dense Block (RDB).

medical imaging applications where clarity and detail are paramount.

The integration of 'Saliency-Aware' mechanisms within the SNAG and CVIAB specifically tailors the model to emphasize salient features in the images, such as critical anatomical structures and surgical instruments. This saliency-aware approach ensures that the most information-rich parts of the image receive higher priority during the SR process, leading to more precise and clinically useful images. The model's ability to discern and enhance these salient features directly supports the enhanced diagnostic and surgical accuracy, which is the primary focus of our research.

$$I_k^{(left,SR)}, I_k^{(right,SR)} = \mathcal{J}\left(\{I_{(k-m)}^{(left,LR)}, \ldots, I_{(k+m)}^{(left,LR)}\}, \{I_{(k-m)}^{(right,LR)}, \ldots, I_{(k+m)}^{(right,LR)}\}, \phi\right) \quad (1)$$

A. FEATURE EXTRACTION

1) StereoNet ATTENTION GATE (SNAG)

The StereoNet Attention Gate (SNAG), as an essential feature extraction mechanism, plays a pivotal role in enhancing the quality of endoscopic images through SR.

Channel attention mechanisms have been widely employed in SR tasks to recalibrate feature responses across the network globally [5], enhancing overall feature sensitivity and improving reconstruction quality. However, these mechanisms typically operate on a global scale, adjusting channel-wise features uniformly, which may not be optimal for specific applications requiring localized attention to detail, such as medical imaging.

In contrast, the StereoNet Attention Gate (SNAG) in our model introduces several innovative enhancements over traditional channel attention, making it particularly well-suited for the nuanced requirements of endoscopic image SR. Firstly, SNAG employs Instance Normalization at the outset, standardizing features across each channel in response to the fluctuating lighting conditions typical in endoscopic procedures. This normalization process ensures more stable network performance and better convergence, a crucial advantage in medical applications where consistency in image quality is paramount.

Following normalization, SNAG utilizes depth-wise separable convolutions that process data more efficiently than the convolutions typically employed in channel attention mechanisms. This efficiency is vital for maintaining the extraction of intricate spatial details necessary for high-quality SR. Moreover, the integration of a Simple Gate mechanism further refines the feature flow within the network. Unlike traditional channel attention that broadly adjusts features across all channels, SNAG's gating mechanism focuses more

selectively on significant features, effectively diminishing the influence of less pertinent ones. This targeted feature refinement allows SNAG to enhance the reconstruction of fine details such as tissue textures and surgical instruments, which are critical for diagnostic accuracy and surgical precision in endoscopic imaging, as depicted in Fig 3(a).

Overall, the SNAG mechanism offers a more adaptable and efficient approach to handling the specific challenges of endoscopic image SR. By focusing on localized and dynamic feature recalibration, SNAG not only outperforms traditional channel attention mechanisms in terms of detail recovery and adaptation to variable imaging conditions but also demonstrates substantial improvements in computational efficiency, making it highly suitable for real-time medical applications. These innovations highlight the distinctive contributions of our model to advancing SR technology, particularly in critical medical imaging contexts.

$$F_{\text{ini}}^{(k)} = \text{Conv}1\times1(\text{SP}(\text{CDA}(\text{SG}(\text{DSC}(\text{IN}(I_k^{(\text{LR})}))))) + I_k^{(\text{LR})} \tag{2}$$

$$F_{\text{SNAG}}^{(k)} = \text{DSC}(\text{SG}(\text{DSC}(\text{IN}(F_{\text{ini}}^{(k)})))) + F_{\text{ini}}^{(k)} \tag{3}$$

2) FEATURE REFINEMENT BLOCK

The Residual Dense Block (RDB) [53] is utilized for further feature refinement after the SNAG block's initial extraction. It is unique to process endoscopic images in a SR context. The RDB consists of several convolutional layers followed by ReLU activations, with shortcut connections that allow for feature reuse and avoid the vanishing gradient problem, which is critical when training deeper networks for complex tasks such as medical image SR. Fig 3 (d).

$$F_{i,\alpha} = f_{\text{RDB}}(F_{SNAG,\alpha}^{(k)}) \quad \text{for } \alpha \in \{\text{left}, \text{right}\} \tag{4}$$

3) CROSS-VIEW INTERACTIVE ATTENTION BLOCK (CVIAB)

The Cross-View Interactive Attention Block (CVIAB), as depicted in Fig 3(b), marks a significant advancement in the domain of stereo endoscopic image SR. It achieves this by promoting a sophisticated interaction between features extracted from the left and right views of endoscopic imagery. At the heart of CVIAB lies the Stereo Attentional Residual Module (SARM), illustrated in Fig 3(c), which is ingeniously designed to refine spatial features through an efficient convolutional strategy. This strategy comprises a blend of depthwise separable and dilated convolutions, tailored specifically for endoscopic image analysis where detailed texture and context capture are paramount. Where DC is dilated convolution

The functional essence of SARM within the CVIAB framework can be encapsulated by the following equation:

$$F_{\text{SARM}}^{(v)} = \text{DC}\left(\text{DSC}\left(F_{\text{RDB}}^{(v)}\right)\right), \quad v \in \{\text{left}, \text{right}\} \tag{5}$$

This expression delineates the sequential processing of features originating from the Residual Dense Block (RDB), whereby the Depthwise Separable Convolution operates first, ensuring computational efficiency while preserving the capability for detailed spatial analysis. Following this, the Dilated Convolution extends the receptive field, enabling a comprehensive contextual grasp without augmenting the computational demand— a critical consideration for the nuanced textural and structural complexity inherent in endoscopic images.

To further elucidate the process of parallax attention and occlusion handling within the CVIAB, consider the stereo images I_L and I_R in the set $\mathbb{R}^{H\times W}$. Parallax attention vectors, $P_{R\to L}$ and $P_{L\to R}$, are derived to spotlight occlusions effectively, thereby transforming I_R into the perspective of I_L as follows:

$$I_{R\to L}(y,:) = P_{R\to L}(y,:,:)\cdot I_R(y,:) \tag{6}$$

The mechanism for occlusion influence mitigation and ambient noise compensation involves computing valid masks M_L and M_R, which are derived through a scaled hyperbolic tangent function, enhancing clarity and interpretability in the super-resolved images:

$$M_v = \tanh\left(\eta \cdot C_v'\right), \quad v \in \{\text{L}, \text{R}\} \tag{7}$$

The integration and fusion of dual-view information in CVIAB are achieved via a unique dual-view attention mechanism, facilitated by depth-wise separable convolutions. This mechanism not only amalgamates but also significantly enriches the features from both views, as depicted in the ensuing formulation:

$$F_{\text{CVIAB}} = U_d^{\text{dir}}\text{Attention}_{\text{dir}}(Z_{\text{i}}, Y_{\text{i}}, X_{\text{i}}), \quad \text{dir} \in \{\text{left}\to\text{right}, \text{right}\to\text{left}\} \tag{8}$$

Through these designed operations, the CVIAB ensures the seamless integration and enhancement of feature representations across both views. This approach not only fortifies the model's capacity for high-fidelity image reconstruction but also significantly contributes to the SR quality, vital for the exigencies of medical diagnostics and procedures in endoscopic imaging.

4) RECONSTRUCTION BLOCK

The Reconstruction Block is pivotal in synthesizing the high-resolution endoscopic image from the extracted features. It integrates a series of operations starting from the Residual Dense Block (RDB) and continuing through channel and depth attention, strip pooling, 1×1 convolution, and depth-wise separable convolution. Each component builds upon the previous one, progressively enhancing the image quality and ensuring the reconstructed image retains the essential details necessary for accurate medical diagnosis.

The entire process within the Reconstruction Block can be concisely represented by the following equation:

$$F_{\text{recon}} = \text{DSC}\Big(\text{Conv}1\times1\Big(\text{SP}\Big(\text{CDA}\Big(F_{\text{CVIAB}}\Big)\Big)\Big)\Big) \tag{9}$$

Here, F_{RDB} denotes the output feature map from the Residual Dense Block, and F_{recon} is the final feature map

output by the Reconstruction Block, ready to be transformed into the super-resolved image.

IV. EXPERIMENTAL RESULTS

A. EXPERIMENTAL SETTINGS

Our training dataset comprised 894 stereo image pairs from the SCARED collection. High-resolution (HR) frames were downscaled using bicubic interpolation to generate low-resolution (LR) counterparts for training. Data augmentation strategies, including vertical image flipping, were applied to enrich the training dataset. For evaluation, we utilized five distinct stereo endoscopic image datasets: the da Vinci dataset, other datasets which can be accessed at this link (https://endovis.grand-challenge.org/), the SCARED dataset, the MICCAI 2017 Kidney Boundary Detection SubChallenge; the Kidney Boundary Detection dataset, the MICCAI 2017 Robotic Instrument Segmentation Sub-Challenge; Robotic Instrument Segmentation, and the MICCAI 2019 challenge on Stereo Correspondence and Reconstruction of Endoscopic Data; Stereo Correspondence and Reconstruction, covering a range of clinical scenarios. This comprehensive testing framework was designed to rigorously assess the performance and adaptability of our model across various clinical conditions.

The model was developed using PyTorch and trained on an NVIDIA 3090ti GPU for computational efficiency. We adopted the Adam optimizer for model optimization, setting $\beta_1 = 0.9$ and $\beta_2 = 0.999$, with a training batch size 8. The learning rate was initiated at 1×10^{-4}. For our experiments, the 'k' parameter was set to 1, indicating using three consecutive frames as input during training.

B. LOSS FUNCTION

We have implemented the same loss function as was used in previous research. The loss function consists of two main parts for a stereo-matching model: the SR loss L_{SR}, and the parallax-attention loss L_{PAM}. The SR loss calculated the similarity between the predicted and HR ground-truth stereo images. In contrast, the parallax-attention loss encouraged the model to emphasize the most salient features of the scene. The stereo consistency loss ensures that the predicted depth maps are consistent with the stereo images.

$$\text{Loss} = L_{SR} + \lambda \left(L_{\text{photometric}} + L_{\text{cycle}} + L_{\text{smoothness}}\right) \quad (10)$$

1) SR LOSS

An $L1$ loss is commonly used to compare the predicted SR stereo images with the corresponding ground-truth stereo images to calculate the SR loss in a stereo-matching model. The L1 loss measures the difference between the pixel values of the two images and is often used over other loss functions due to its robustness to outliers.

$$L_{SR} = \left\| I_{\text{left}}^{SR} - I_{\text{left}}^{HR} \right\|_1 + \left\| I_{\text{right}}^{SR} - I_{\text{right}}^{HR} \right\|_1 \quad (11)$$

I_{left}^{SR} and I_{left}^{HR} represent output SR and HR images respectively, and similarly for right view images.

2) PARALLAX-ATTENTION LOSS

The parallax-attention loss in a stereo-matching model typically comprises three terms: photometric, smoothness, and cycle. These terms are combined to result in a loss function that encourages the model to address the most relevant features in the scene while maintaining smoothness and consistency across the predicted depth maps. The photometric term ensures that the predicted stereo images are consistent with the input images, while the smoothness term encourages smooth transitions between neighboring depth values. The cycle term ensures that the predicted depth maps can be used to reconstruct the input images.

$$L_{PAM} = \lambda \left(L_{\text{photometric}} + L_{\text{cycle}} + L_{\text{smoothness}}\right) \quad (12)$$

$$\begin{aligned} L_{\text{photometric}} &= \sum_{\text{pixel} \in V_{\text{left-right}}} \left| I_{\text{left}}^{LR}(\text{pixel}) - (M_{\text{left-right}} * I_{\text{right}}^{LR})(\text{pixel}) \right|_1 \\ &+ \sum_{\text{pixel} \in V_{\text{right-left}}} \left| I_{\text{right}}^{LR}(\text{pixel}) - (M_{\text{right-left}} * I_{\text{left}}^{LR})(\text{pixel}) \right|_1 \end{aligned} \quad (13)$$

$$\begin{aligned} L_{\text{cycle}} &= \sum_{\text{pixel} \in V_{\text{left-right}}} \left| (M_{\text{left-right-left}}(\text{pixel}) - I_{\text{pixel}}) \right|_1 \\ &+ \sum_{\text{pixel} \in V_{\text{right-left}}} \left| (M_{\text{right-left-right}}(\text{pixel}) - I_{\text{pixel}}) \right|_1 \end{aligned} \quad (14)$$

where $I \in \mathbb{R}^{H \times W \times W}$ and M includes $M_{\text{left-right}}$, $M_{\text{right-left}}$.

$$\begin{aligned} L_{\text{smoothness}} = \sum_M \Bigg[\sum_{i,j,k} &|(M(i,j,k) - M(i+1,j,k))|_1 \\ &+ |(M(i,j,k) - M(i,j+1,k))|_1 \Bigg] \end{aligned} \quad (15)$$

C. EVALUATION RESULTS

We evaluated image SR performance using peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM), essential for assessing the quality and similarity of high-resolution (HR) and super-resolved images. These metrics were calculated within the RGB color space and averaged for both left and right images of stereo pairs.

Our proposed network demonstrated superior PSNR and SSIM scores for both $\times 2$, $\times 4$, and $\times 8$ SR tasks across all datasets, outperforming existing single, stereo, and video SR methods as shown in Table 1. This indicates our model's effectiveness in employing temporal cross-attention and parallel attention mechanisms for high-quality image reconstruction.

Qualitative evaluations through zoomed-in comparisons on the SCARED, Robotic Instrument Segmentation, and Stereo Correspondence and Reconstruction datasets presented in Fig. 4-12 further illustrate our model's ability to capture fine details more accurately than single-image

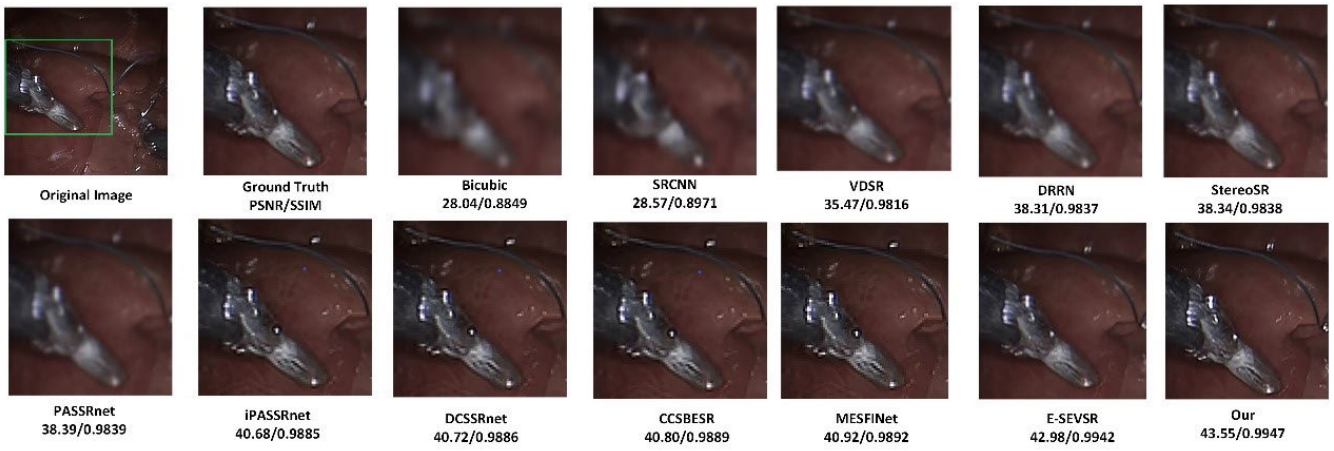

FIGURE 4. Assessing the perceptual quality of SR images with a ×2 scale on the da vinci dataset.

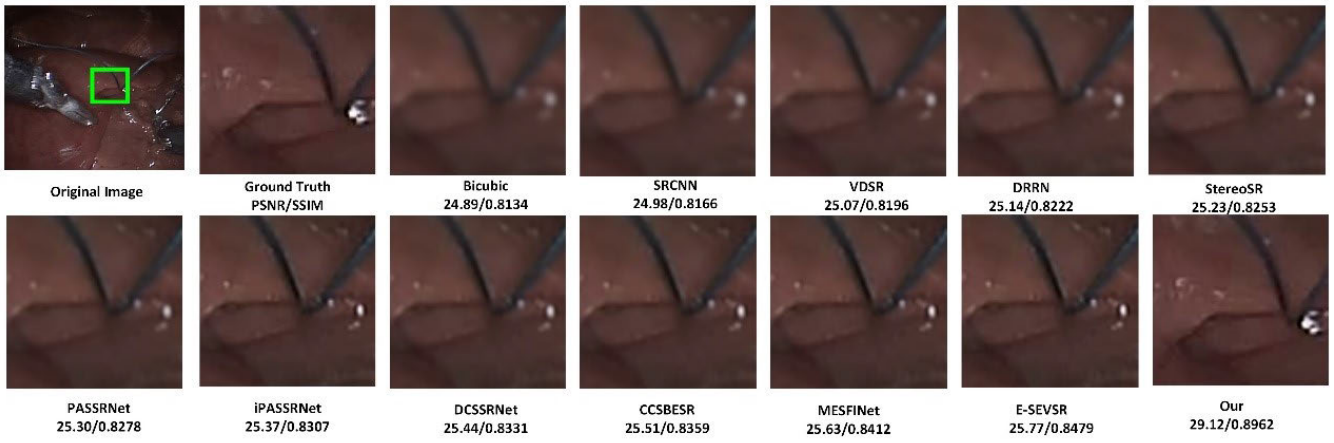

FIGURE 5. Assessing the perceptual quality of SR images with a ×8 scale on the da vinci dataset.

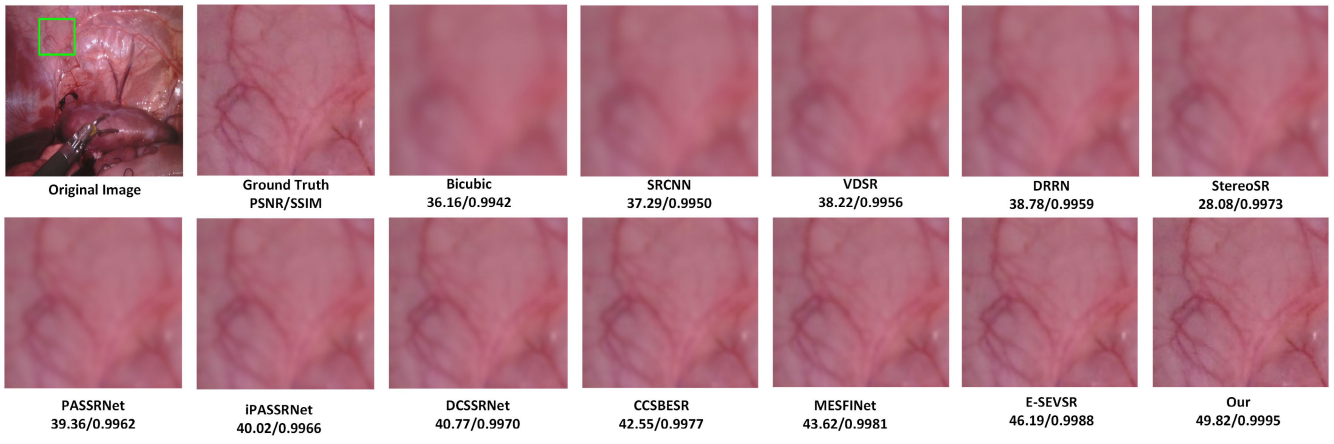

FIGURE 6. Assessing the perceptual quality of SR images with a ×2 scale factor on the scared dataset.

SR methods, resulting in more transparent and higher-quality images. This underscores our model's robustness in complex scenarios, marking a significant improvement in SR technology.

1) EVALUATION RESULTS ON REAL- WORLD DATA

To ensure the robustness and applicability of our proposed model, we meticulously evaluated its performance on two real-world image datasets: Flickr1024 [54] and

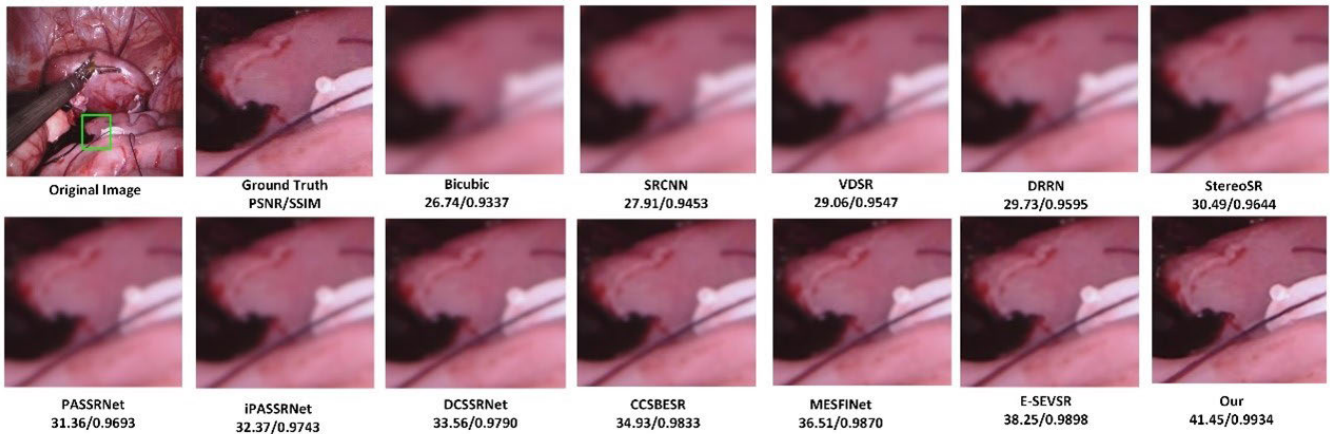

FIGURE 7. Assessing the perceptual quality of SR images with a ×4 scale factor on the scared dataset.

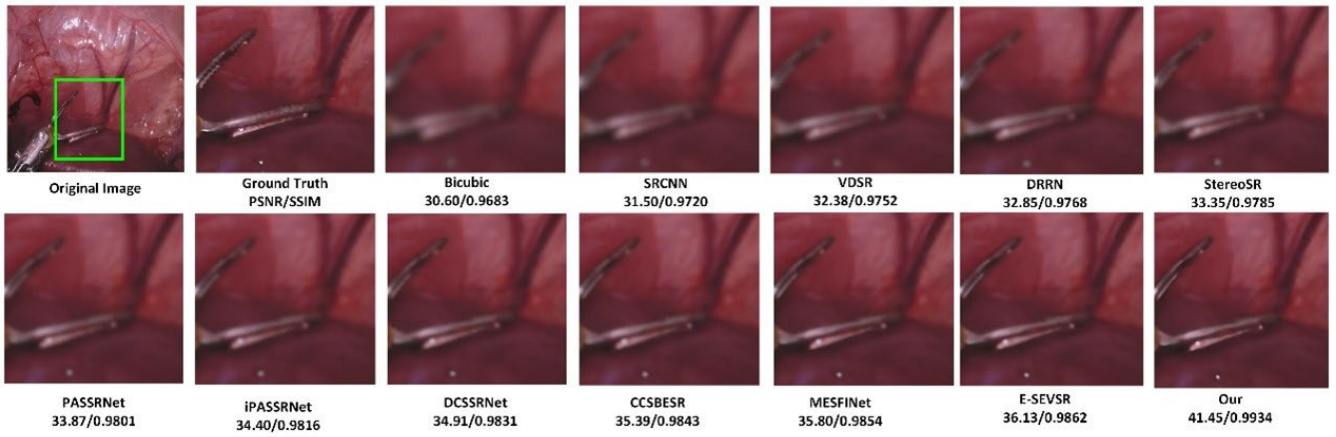

FIGURE 8. Assessing the perceptual quality of SR images with a ×8 scale factor on the scared dataset.

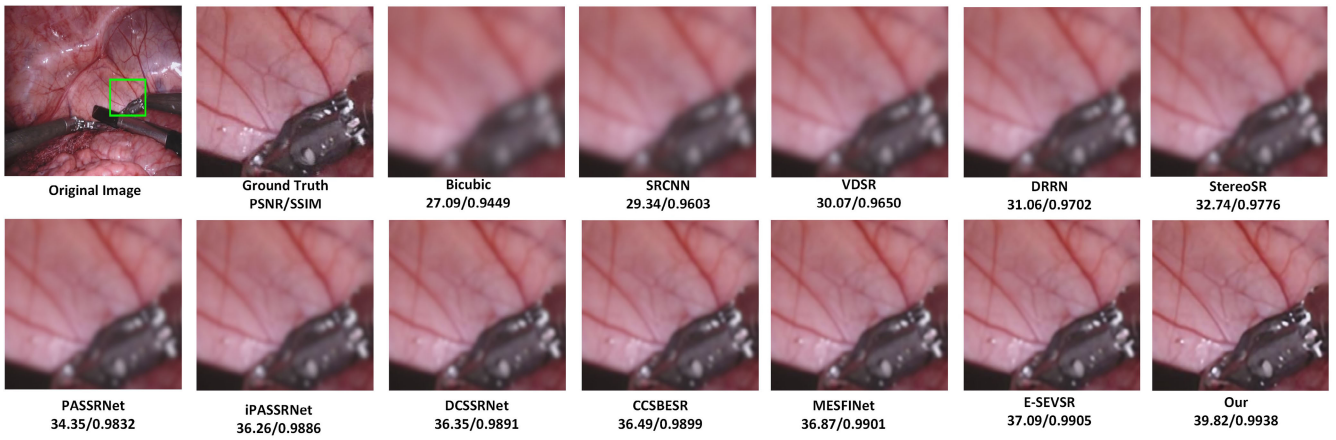

FIGURE 9. Assessing the perceptual quality of SR images with a ×4 scale factor on the robotic instrument segmentation dataset.

Middlebury [55]. These datasets were selected due to their diverse and challenging nature, which includes a variety of scenes and image conditions.

Our model demonstrated impressive performance on both datasets, producing high-quality visual results that are illustrated in Fig. 13-15. These figures showcase our

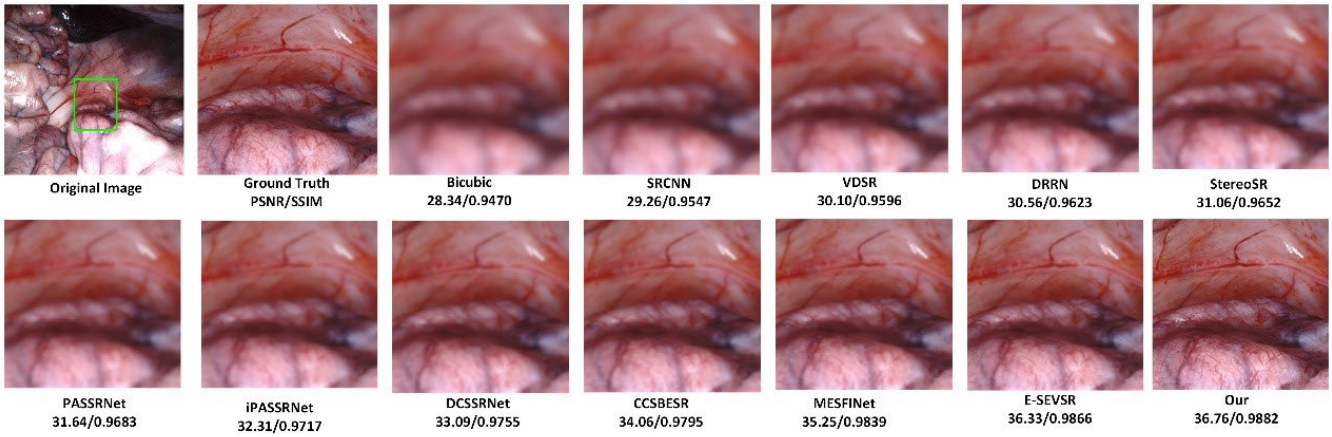

FIGURE 10. Assessing the perceptual quality of SR images with a ×2 scale factor on the stereo correspondence and reconstruction dataset.

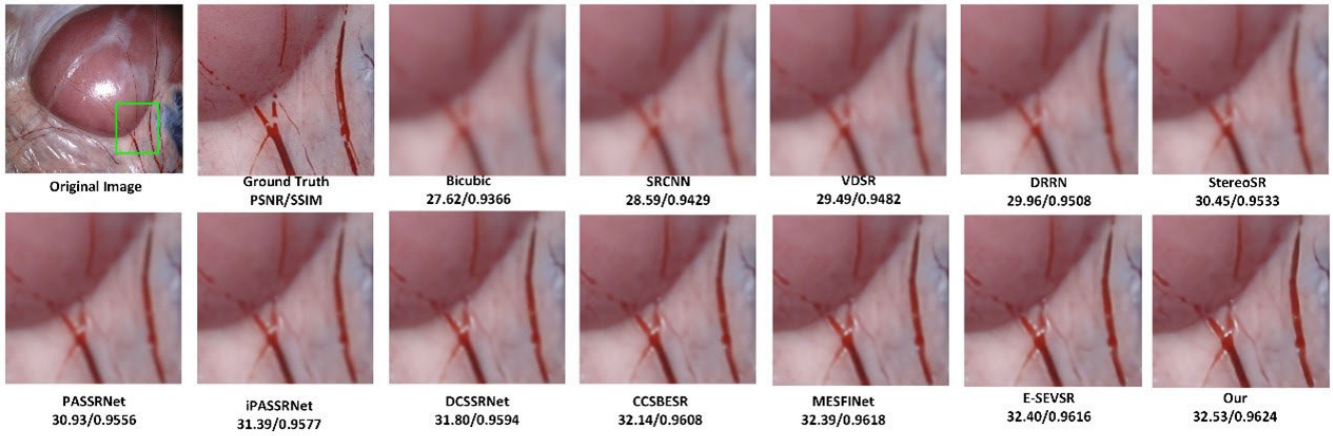

FIGURE 11. AAssessing the perceptual quality of SR images with a ×4 scale factor on the stereo correspondence and reconstruction dataset.

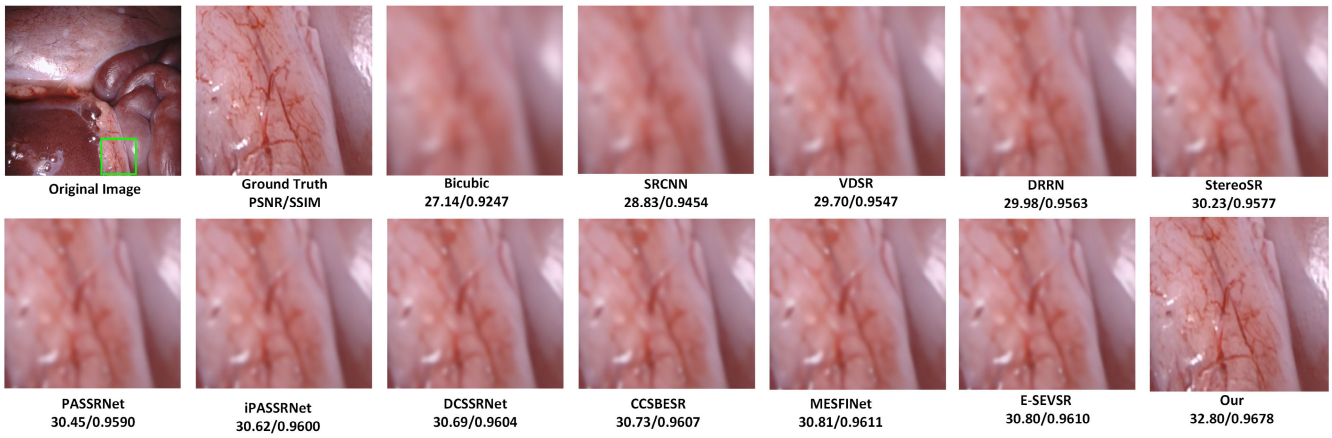

FIGURE 12. Assessing the perceptual quality of SR images with a ×8 scale factor on the stereo correspondence and reconstruction dataset.

model's enhanced image details and clarity, validating its effectiveness in real-world scenarios. The visual results highlight the model's ability to preserve fine details and improve image resolution significantly, making it a valuable tool for practical applications.

Model Efficiency Analysis: Parameters, FLOPs and Inference Time: This analysis compares various SR models, highlighting the efficiency of computational demand (measured in Giga FLOPs) and model compactness (measured in millions of parameters). A noteworthy point in modern computational

TABLE 1. Quantitative comparison using PSNR/SSIM on da vinci dataset, scared, kidney boundary detection, robotic instrument segmentation and stereo correspondence and reconstruction with enlargement factor ×2, ×4 and ×8.

Method	Scale	da Vinci	SCARED	Kidney Boundary Detection	Robotic Instrument Segmentation	Stereo Correspondence and Reconstruction
bicubic [38]	×2	32.12/0.9799	32.26/0.9791	33.39/0.9736	31.91/0.9701	31.95/0.9755
SRCNN [42]	×2	36.18/0.9868	36.10/0.9890	37.55/0.9873	35.01/0.9815	35.10/0.9801
VDSR [21]	×2	40.71/0.9906	39.82/0.9931	38.01/0.9887	35.78/0.9830	35.91/0.9813
DRRN [39]	×2	40.80/0.9907	39.91/0.9933	38.02/0.9889	35.75/0.9832	35.95/0.9815
StereoSR [43]	×2	40.95/0.9909	39.99/0.9936	38.10/0.9890	35.91/0.9834	36.01/0.9817
PASSRNet [27]	×2	41.89/0.9919	40.95/0.9941	38.99/0.9899	36.18/0.9839	36.75/0.9825
iPASSRNet [9]	×2	41.94/0.9921	41.01/0.9941	39.01/0.9901	36.22/0.9841	36.79/0.9827
DCSSRNet [10]	×2	42.79/0.9940	41.74/0.9955	39.85/0.9912	36.89/0.9875	36.94/0.9830
CCSBESR [5]	×2	42.81/0.9942	41.76/0.9960	39.89/0.9914	36.91/0.9877	36.97/0.9832
MESFINet [11]	×2	42.89/0.9943	41.79/0.9961	39.91/0.9915	36.94/0.9878	36.99/0.9834
Trans-SVSR [50]	×2	38.21/0.9767	40.73/0.9875	36.31/0.9896	36.96/0.9878	37.00/0.9835
HA-VSR [40]	×2	38.37/0.9771	40.80/0.9870	39.81/0.9913	36.99/0.9879	37.01/0.9835
E-SEVSR [6]	×2	43.01/0.9945	41.84/0.9963	40.01/0.9916	37.01/0.9880	37.02/0.9836
Our	×2	**43.64/0.9947**	**42.30/0.9965**	**40.51/0.9920**	**37.70/0.9884**	**37.96/0.9841**
bicubic [38]	×4	28.24/0.9489	33.69/0.9594	27.69/0.9389	32.79/0.9376	31.99/0.9381
SRCNN [42]	×4	29.11/0.9501	34.59/0.9661	29.49/0.9440	33.59/0.9420	32.69/0.9434
VDSR [21]	×4	30.89/0.9599	35.41/0.9704	32.51/0.9501	34.03/0.9497	33.57/0.9528
DRRN [39]	×4	30.94/0.9600	35.49/0.9704	32.54/0.9503	34.06/0.9499	33.61/0.9531
StereoSR [43]	×4	31.01/0.9601	35.59/0.9705	32.69/0.9505	34.16/0.9500	33.81/0.9533
PASSRNet [27]	×4	31.31/0.9603	35.79/0.9707	32.76/0.9515	34.25/0.9506	33.99/0.9538
iPASSRNet [9]	×4	31.34/0.9603	35.81/0.9708	32.79/0.9546	34.29/0.9508	34.01/0.9539
DCSSRNet [10]	×4	34.41/0.9605	35.94/0.9709	32.92/0.9548	34.41/0.9511	34.02/0.9539
CCSBESR [5]	×4	31.53/0.9607	36.01/0.9710	33.02/0.9551	34.49/0.9512	34.02/0.9540
MESFINet [11]	×4	31.61/0.9609	36.09/0.9711	33.09/0.9551	34.64/0.9514	34.06/0.9540
Trans-SVSR [50]	×4	31.81/0.9469	34.68/0.9573	27.33/0.9403	34.71/0.9514	34.10/0.9540
HA-VSR [40]	×4	32.03/0.9477	34.79/0.9576	28.59/0.9585	34.68/0.9515	34.14/0.9541
E-SEVSR [6]	×4	32.01/0.9615	36.99/0.9718	33.13/0.9551	34.84/0.9515	34.19/0.9541
Our	×4	**33.98/0.9734**	**38.43/0.9926**	**35.11/0.9763**	**36.74/0.9829**	**36.12/0.9749**
bicubic [38]	×8	23.01/0.9211	26.01/0.9501	28.01/0.9509	28.56/0.9578	28.09/0.9481
SRCNN [42]	×8	27.63/0.9281	29.79/0.9696	30.01/0.9583	31.51/0.9668	30.11/0.9494
VDSR [21]	×8	27.68/0.9283	29.89/0.9701	30.15/0.9586	31.84/0.9672	30.51/0.9500
DRRN [39]	×8	27.71/0.9284	29.99/0.9710	30.21/0.9591	31.97/0.9674	30.63/0.9501
StereoSR [43]	×8	27.76/0.9285	30.30/0.9746	30.73/0.9594	32.06/0.9676	30.71/0.9503
PASSRNet [27]	×8	27.84/0.9288	30.84/0.9752	31.00/0.9598	32.30/0.9681	30.86/0.9506
iPASSRNet [9]	×8	27.86/0.9288	30.85/0.9752	31.03/0.9600	32.31/0.9681	30.87/0.9507
DCSSRNet [10]	×8	27.90/0.9291	30.86/0.9752	31.11/0.9602	32.33/0.9683	30.90/0.9507
CCSBESR [5]	×8	27.97/0.9294	30.95/0.9754	31.13/0.9604	32.36/0.9685	30.99/0.9509
MESFINet [11]	×8	28.01/0.9296	31.09/0.9756	31.21/0.9605	32.44/0.9686	31.01/0.9511
Trans-SVSR [50]	×8	28.06/0.9297	31.13/0.9756	31.29/0.9606	32.48/0.9686	31.03/0.9511
HA-VSR [40]	×8	28.11/0.9298	31.16/0.9757	31.33/0.9606	32.51/0.9687	31.05/0.9511
E-SEVSR [6]	×8	28.17/0.9299	31.19/0.9757	31.41/0.9607	32.66/0.9688	31.06/0.9511
Our	×8	**29.21/0.9396**	**33.20/0.9811**	**32.08/0.9631**	**32.74/0.9691**	**32.03/0.9533**

models, especially transformer-based ones, is their substantial computational expense, which can limit their practical deployment in resource-constrained environments. Table 2.

The data indicates that transformer-based models, such as Trans-SVSR [50] and HA-VSR [40], are particularly resource-intensive, often requiring significantly higher FLOPs. In contrast, our model achieves average FLOPs (9.91 G) compared to these models. This efficiency not only suggests an optimal balance between computational cost and performance but also positions as a leading choice for practical SR applications, especially in environments with limited computational resources.

The inference times for various SR models reflect the trade-off between model complexity and performance. Simpler models like Bicubic interpolation and SRCNN have very low inference times of 0.46 ms and 0.7 ms, respectively. In contrast, more complex models such as DRRN and advanced stereo models like StereoSR and iPASSRNet exhibit longer inference times of 300 ms, 100.10 ms, and 186.97 ms, respectively. Notably, transformer-based models

FIGURE 13. Scene 17 Left View (Flickr1024) on ×4.

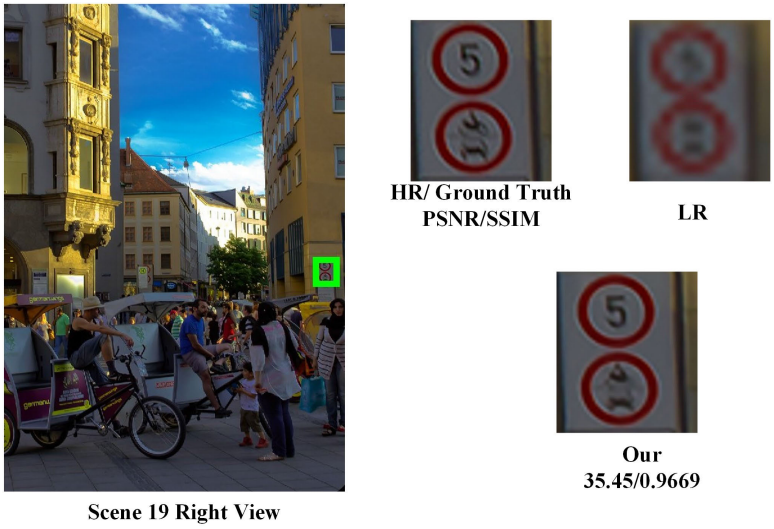

FIGURE 14. Scene 19 Right View (Flickr1024) on ×4.

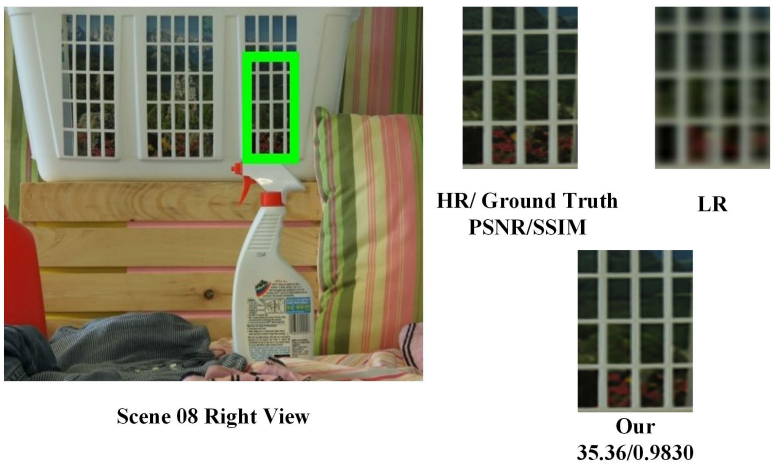

FIGURE 15. Scene 08 Right View (Middlebury) on ×4.

like Trans-SVSR and HA-VSR have significantly higher inference times of 448.6 ms and 432.53 ms due to their extensive network architectures and attention mechanisms. With an inference time of 199.47 ms, our model strikes a balance by offering high performance with comparatively lower latency, making it more suitable for real-time applications.

TABLE 2. Comparison of parameters, FLOPs, interference time (on ×2) for SR models.

Method	#Params. (M)	#FLOPs (G)	Inference Time (ms)
Bicubic [38]	-	-	0.46
SRCNN [42]	0.07	52.7	0.7
VDSR [21]	0.64	10.77	2.0
DRRN [39]	0.29	6.79	300
StereoSR [43]	1.08	90.11	100.10
PASSRNet [27]	1.37	6.53	176
iPASSRNet [9]	1.38	7.44	186.97
DCSSRnet [10]	1.37	6.94	200.19
CCSBESR [5]	1.76	6.96	205.67
MESFINet [11]	2.98	11.76	248.97
Trans-SVSR [50]	27.29	86.54	448.6
HA-VSR [40]	42.54	99.54	432.53
E-SEVSR [6]	2.99	11.80	252.81
OUR	2.84	9.91	199.47

TABLE 3. Ablation study integrating different FE Block using scared dataset on ×2.

	Conv	CCSB	CDCA	SNAG	PSNR/SSIM
FE Blocks	✓	×	×	×	38.41/0.9921
	✓	✓	×	×	40.50/0.9931
	✓	×	✓	×	40.54/0.9931
	×	×	×	✓	**42.30/0.9965**

V. ABLATION STUDY

1) FEATURE EXTRACTION

An ablation study was performed using the SCARED dataset at a x2 upscaling to evaluate the performance of various feature extraction blocks. The study compared traditional convolution (Conv), Combined Channel and Spatial Attention (CCSB), Combined Depth and Channel Attention (CDCA), and the novel StereoNet Attention Gate (SNAG). Results indicated that CDCA and SNAG blocks significantly enhance image quality, with SNAG achieving the highest PSNR/SSIM scores of 42.30/0.9965, demonstrating its superior feature extraction capability. Table 3

2) CROSS-VIEW FEATURE INTERACTION

An ablation study on the SCARED dataset with x2 upscaling assessed different cross-view feature interaction blocks: Stereo Cross-Attention Module (SCAM), bidirectional Position Attention Module (biPAM), and the novel Cross-View Interactive Attention Block (CVIAB). The CVIAB block, notably designed for this research, achieved superior image enhancement, evidenced by the highest PSNR/SSIM of 42.30/0.9965, validating its efficacy in cross-view feature extraction. Table 4

VI. LIMITATIONS AND FUTURE WORK

Our research establishes a significant benchmark in the Stereo Endoscopic Image SR domain, meticulously adhering to the experimental frameworks established by prior studies. While our model is adept at stereo endoscopic image enhancement, it may not perform comparably on real-world scene super-resolution, as it is not specifically optimized for such applications. Despite its strengths, our current model

TABLE 4. Ablation study integrating different cross view feature interaction blocks using scared dataset on ×2.

	SCAM	biPAM	CVIAB	PSNR/SSIM
Cross View Feature Interaction	✓	×	×	38.01/0.9891
	×	✓	×	38.29/0.9900
	×	×	✓	**42.30/0.9965**

falls short of supporting real-time SR in endoscopic surgeries, primarily due to computational bottlenecks and a shortage of specialized resources required for immediate application in surgical scenarios. To transcend these limitations, future iterations of our model could benefit from the integration of advanced features such as motion estimation modules, frame interpolation techniques, and feature temporal interpolation strategies. Furthermore, embracing hardware innovations—such as deploying an array of GPUs for enhanced parallel processing, utilizing high-speed Input/Output interfaces, incorporating Field-Programmable Gate Arrays (FPGA), exploring server clusters, or developing Application-Specific Integrated Circuits (ASIC)—could significantly elevate the model's real-time processing prowess.

Expanding beyond its initial scope, our model holds the potential for adaptation across a wider spectrum of medical imaging technologies, including but not limited to MRI, CT, and PET scans. This prospective broadening of application could revolutionize not only the field of endoscopic surgery but also the broader realm of medical diagnostics and treatment planning, offering improved resolution and clarity in imaging across various modalities. Such advancements and modifications are not only technically feasible but also strategically align with the overarching goal of enabling our model's practical deployment in real-time surgical environments and potentially transforming clinical practices at large. These directions not only aim to surmount the current operational constraints but also aspire to significantly extend the utility, performance, and impact of our model in diverse clinical settings.

VII. CONCLUSION

This research represents a significant advancement in the field of endoscopic image SR through the introduction of the StereoNet Attention Gate (SNAG) and CVIAB, its integration within our SR framework. Our model has demonstrated a remarkable ability to enhance image resolution, particularly at high scales, which is pivotal for improving the clarity and utility of endoscopic images in medical diagnostics and surgical planning.

While our approach has set new benchmarks in the accuracy and quality of super-resolved images, it is not without limitations. One of the primary challenges lies in the real-time application of the technology in clinical settings. Although optimized, the computational demands of our current model still require substantial resources that may only be readily available in some medical facilities. Looking to the future, several enhancements are possible to

overcome these limitations. First, further optimization of the computational architecture is necessary to facilitate real- time processing capabilities. This may involve integrating more efficient convolutional operations or adopting newer, faster processing units. Moreover, extending the model's robustness to handle diverse and unforeseen surgical environments by training with a broader range of data scenarios will be significant. Another promising direction is the exploration of transfer learning techniques to adapt the model for different types of endoscopic procedures without extensive retraining. In addition, exploring the integration of motion estimation and correction algorithms to enhance the model's applicability in dynamic surgical scenarios, potentially increasing its utility in a wider range of medical procedures, can be a potential future direction. These advancements not only aim to address the current operational constraints but also broaden our model's impact and applicability in the medical field, thereby transforming clinical practices and improving patient outcomes.

In conclusion, our study, while highlighting significant achievements, also acknowledges the inherent challenges and limitations faced by current SR technologies in medical applications. By addressing these challenges head-on and setting a clear path for future research, we hope to pave the way for more robust, efficient, and widely applicable SR solutions in medical imaging.

DATA AVAILABILITY STATEMENT

The datasets analyzed during the current study are from the following publicly available sources: The SCARED, the MICCAI 2017 Kidney Boundary Detection Sub-Challenge, the MICCAI 2017 Robotic Instrument Segmentation Sub-Challenge, the MICCAI 2019 Challenge on Stereo Correspondence and Reconstruction of Endoscopic Data and the EndoVis datasets accessible at (https://endovis.grand-challenge.org/). An additional da Vinci dataset is available at (https://github.com/hgfe/DCSSR).

LIST OF ABBREVIATIONS

TABLE 5. Abbreviations.

Abbreviation	Full Form
SNAG	StereoNet Attention Gate
RDB	Residual Dense Block
CVIAB	Cross-View Interactive Attention Block
SARM	Stereo Attentional Residual Module
PSNR	Peak Signal-to-Noise Ratio
SSIM	Structural Similarity Index Measure
SR	Super-Resolution
CNN	Convolutional Neural Network
GAN	Generative Adversarial Network
RNN	Recurrent Neural Network
SP	Strip Pooling
CDA	Channel & Depth Attention
SG	Simple Gate
DSC	Depthwise Seperable Convolution
IN	Instance Normalization
DC	Dilated Convolution

REFERENCES

[1] B. S. Peters, P. R. Armijo, C. Krause, S. A. Choudhury, and D. Oleynikov, "Review of emerging surgical robotic technology," *Surgical Endoscopy*, vol. 32, no. 4, pp. 1636–1655, Apr. 2018.

[2] U. D. A. Mueller-Richter, A. Limberger, P. Weber, K. W. Ruprecht, W. Spitzer, and M. Schilling, "Possibilities and limitations of current stereo-endoscopy," *Surgical Endoscopy*, vol. 18, no. 6, pp. 942–947, Jun. 2004.

[3] C.-C. Wang, Y.-C. Chiu, W.-L. Chen, T.-W. Yang, M.-C. Tsai, and M.-H. Tseng, "A deep learning model for classification of endoscopic gastroesophageal reflux disease," *Int. J. Environ. Res. Public Health*, vol. 18, no. 5, p. 2428, Mar. 2021.

[4] S. Ali et al., "Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy," *Med. Image Anal.*, vol. 70, May 2021, Art. no. 102002.

[5] M. Hayat, S. Aramvith, and T. Achakulvisut, "Combined channel and spatial attention-based stereo endoscopic image super-resolution," in *Proc. TENCON IEEE Region 10 Conf. (TENCON)*, Chiang Mai, Thailand, Oct. 2023, pp. 920–925, doi: 10.1109/TENCON58879.2023.10322331.

[6] M. Hayat and S. Aramvith, "E-SEVSR—Edge guided stereo endoscopic video super-resolution," *IEEE Access*, vol. 12, pp. 30893–30906, 2024, doi: 10.1109/access.2024.3367980.

[7] B. Yan, C. Ma, B. Bare, W. Tan, and S. Hoi, "Disparity-aware domain adaptation in stereo image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13176–13184.

[8] D. S. Jeon, S.-H. Baek, I. Choi, and M. H. Kim, "Enhancing the spatial resolution of stereo images using a parallax prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Sep. 2018, pp. 1721–1730.

[9] Y. Wang, X. Ying, L. Wang, J. Yang, W. An, and Y. Guo, "Symmetric parallax attention for stereo image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 766–775.

[10] T. Zhang, Y. Gu, X. Huang, J. Yang, and G.-Z. Yang, "Disparity-constrained stereo endoscopic image super-resolution," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 17, no. 5, pp. 867–875, May 2022.

[11] J. Wan, H. Yin, Z. Liu, Y. Liu, and S. Wang, "Multi-stage edge-guided stereo feature interaction network for stereoscopic image super-resolution," *IEEE Trans. Broadcast.*, vol. 69, no. 2, pp. 357–368, Jun. 2023.

[12] X. Wang, K. C. K. Chan, K. Yu, C. Dong, and C. C. Loy, "EDVR: Video restoration with enhanced deformable convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, p. 1.

[13] W.-C. Siu and K.-W. Hung, "Review of image interpolation and SR," in *Proc. Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, Aug. 2012, pp. 1–10.

[14] A. V. Bhavsar and A. N. Rajagopalan, "Resolution enhancement in multi-image stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1721–1728, Sep. 2010.

[15] Y. Chang, "Research on de-motion blur image processing based on deep learning," *J. Vis. Commun. Image Represent.*, vol. 60, pp. 371–379, Apr. 2019.

[16] H. Hu, S. Yang, X. Li, Z. Cheng, T. Liu, and J. Zhai, "Polarized image SR via a deep convolutional neural network," *Opt. Exp.*, vol. 31, no. 5, pp. 8535–8547, 2023.

[17] K. Zhang, L. V. Gool, and R. Timofte, "Deep unfolding network for image SR," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, Jun. 2020, pp. 3217–3226.

[18] K. Zhang, W. Zuo, and L. Zhang, "Deep plug-and-play super-resolution for arbitrary blur kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1671–1681.

[19] L. Wang, X. Dong, Y. Wang, X. Ying, Z. Lin, W. An, and Y. Guo, "Exploring sparsity in image super-resolution for efficient inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4917–4926.

[20] W. Muhammad, S. Aramvith, and T. Onoye, "SENext: Squeeze-and-ExcitationNext for single image SR," *IEEE Access*, vol. 11, pp. 45989–46003, 2023.

[21] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image SR using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1646–1654.

[22] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 7, pp. 2480–2495, Jul. 2021.

[23] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image SR using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 286–301.

[24] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," 2019, *arXiv:1903.10082*.

[25] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image SR," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 11065–11074.

[26] W. Muhammad, S. Aramvith, and T. Onoye, "Multi-scale xception based depthwise separable convolution for single image super-resolution," *PLoS ONE*, vol. 16, no. 8, Aug. 2021, Art. no. e0249278.

[27] L. Wang, "Learning parallax attention for stereo image SR," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 12250–12259.

[28] Q. Xu, L. Wang, Y. Wang, W. Sheng, and X. Deng, "Deep bilateral learning for stereo image super-resolution," *IEEE Signal Process. Lett.*, vol. 28, pp. 613–617, 2021.

[29] X. Chu, L. Chen, and W. Yu, "NAFSSR: Stereo image SR using NAFNet," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 1239–1248.

[30] C. Chen, Z. Xiong, X. Tian, Z.-J. Zha, and F. Wu, "Camera lens SR," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 1652–1660.

[31] Y. Hang, Q. Liao, W. Yang, Y. Chen, and J. Zhou, "Attention cube network for image restoration," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2562–2570.

[32] Z. Xiong, D. Xu, X. Sun, and F. Wu, "Example-based super-resolution with soft information and decision," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1458–1465, Oct. 2013.

[33] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video SR with convolutional neural networks," *IEEE Trans. Comput. Imag.*, vol. 2, no. 2, pp. 109–122, Mar. 2016.

[34] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia, "Detail-revealing deep video SR," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4472–4480.

[35] L. Pan, Y. Dai, M. Liu, and F. Porikli, "Simultaneous stereo video deblurring and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6987–6996.

[36] A. Sellent, C. Rother, and S. Roth, "Stereo video deblurring," in *Proc. 14th Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands. Cham, Switzerland: Springer, 2016, pp. 558–575.

[37] B. Li, C.-W. Lin, B. Shi, T. Huang, W. Gao, and C.-C. J. Kuo, "Depth-aware stereo video retargeting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6517–6525.

[38] J. Sun, Z. Xu, and H.-Y. Shum, "Image SR using gradient profile prior," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Sep. 2008, pp. 1–8.

[39] Y. Tai, J. Yang, and X. Liu, "Image SR via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Sep. 2017, pp. 3147–3155.

[40] T. Zhang and J. Yang, "Transformer with hybrid attention mechanism for stereo endoscopic video super resolution," *Symmetry*, vol. 15, no. 10, p. 1947, Oct. 2023.

[41] D. Smet, Vincent, V. Namboodiri, and L. Van Gool, "SR techniques for minimally invasive surgery," in *Proc. AE-CAI*, 2011, pp. 41–50.

[42] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image SR," in *Proc. Comput. Vis. ECCV, 13th Eur. Conf.*, Zurich, Switzerland, Cham, Switzerland: Springer, 2014, pp. 184–199.

[43] W. Song, S. Choi, S. Jeong, and K. Sohn, "Stereoscopic image SR with stereo consistent feature," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 12031–12038.

[44] M. Hayat, S. Aramvith, and T. Achakulvisut, "SEGSRNet for stereo-endoscopic image super-resolution and surgical instrument segmentation," 2024, *arXiv:2404.13330*.

[45] K. Jiang, Z. Wang, P. Yi, G. Wang, K. Gu, and J. Jiang, "ATMFN: Adaptive-threshold-based multi-model fusion network for compressed face hallucination," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2734–2747, Oct. 2020.

[46] K. Jiang, Z. Wang, P. Yi, T. Lu, J. Jiang, and Z. Xiong, "Dual-path deep fusion network for face image hallucination," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 1, pp. 378–391, Jan. 2022.

[47] Y. Xiao, Q. Yuan, K. Jiang, J. He, X. Jin, and L. Zhang, "EDiffSR: An efficient diffusion probabilistic model for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5601514.

[48] Y. Xiao, Q. Yuan, K. Jiang, J. He, C.-W. Lin, and L. Zhang, "TTST: A top-k token selective transformer for remote sensing image super-resolution," *IEEE Trans. Image Process.*, vol. 33, pp. 738–752, 2024.

[49] Y. Xiao, Q. Yuan, K. Jiang, X. Jin, J. He, L. Zhang, and C.-W. Lin, "Local-global temporal difference learning for satellite video super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2789–2802, Apr. 2024.

[50] H. Imani, M. B. Islam, and L.-K. Wong, "A new dataset and transformer for stereoscopic video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 706–715.

[51] N. M. Hussain, A. U. Rehman, M. T. B. Othman, J. Zafar, H. Zafar, and H. Hamam, "Accessing artificial intelligence for fetus health status using hybrid deep learning algorithm (AlexNet-SVM) on cardiotocographic data," *Sensors*, vol. 22, no. 14, p. 5103, Jul. 2022.

[52] A. Raza, H. Ayub, J. A. Khan, I. Ahmad, A. S. Salama, Y. I. Daradkeh, D. Javeed, A. Ur Rehman, and H. Hamam, "A hybrid deep learning-based approach for brain tumor classification," *Electronics*, vol. 11, no. 7, p. 1146, Apr. 2022.

[53] Y. Zhang, Y Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image SR," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Sep. 2018, pp. 2472–2481.

[54] Y. Wang, L. Wang, J. Yang, W. An, and Y. Guo, "Flickr1024: A large-scale dataset for stereo image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3852–3857.

[55] D. Scharstein, "High-resolution stereo datasets with subpixel-accurate ground truth," in *Proc. German Conf. Pattern Recognit.*, Sep. 2014, pp. 31–42.

MANSOOR HAYAT (Graduate Student Member, IEEE) received the B.Sc. degree in electrical engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2015, the M.Sc. degree in electrical engineering from the Institute of Southern Multan, Pakistan, in 2018, and the Master of Business Administration (M.B.A.) degree from the National College of Business Administration and Economics, Lahore, Pakistan, in 2022. He is currently pursuing the Ph.D. degree in electrical engineering with Chulalongkorn University, Bangkok, Thailand. His research interests include the application of deep learning and machine learning in medical imaging and video processing. Additionally, he was honored with the Best Conference Paper Award at the 2023 IEEE Region 10 Conference (TENCON), Chiang Mai, Thailand.

SUPAVADEE ARAMVITH (Senior Member, IEEE) received the B.S. degree (Hons.) in computer science from Mahidol University, in 1993, and the M.S. and Ph.D. degrees in electrical engineering from the University of Washington, Seattle, USA, in 1996 and 2001, respectively. She joined Chulalongkorn University, in June 2001, where she is currently an Associate Professor with the Department of Electrical Engineering, specializing in image and video signal processing. She has successfully advised 14 Ph.D., 30 master's, and 41 bachelor's graduates. She has published over 130 papers in international conference proceedings and journals with four international book chapters. She has rich project management experience as the Project Leader and the Former Technical Committee Chair to the Thailand Government bodies in telecommunications and ICT. She is very active in the international arena with leadership positions in the global network, such as JICA Project for AUN/SEED-Net and NICT ASEAN IVO, and professional organizations, such as IEEE, IEICE, APSIPA, and ITU.