IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 February 2024, accepted 15 February 2024, date of publication 20 February 2024, date of current version 1 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3367980

RESEARCH ARTICLE

E-SEVSR–Edge Guided Stereo Endoscopic Video 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 research is funded by Graduate Scholarship Programme for ASEAN or Non–ASEAN Countries, Ratchadapiseksompotch Fund Chulalongkorn University, and the NSRF via the Program Management Unit for Human Resources & Institutional Development, Research and Innovation [grant number B04G640053].

ABSTRACT Integrating Stereo Imaging technology into medical diagnostics and surgeries marks a significant revolution in medical sciences. This advancement gives surgeons and physicians a deeper understanding of patients' organ anatomy. However, like any technology, stereo cameras have their limitations, such as low resolution (LR) and output images that are often blurry. Our paper introduces a novel approach—a multi-stage network with a pioneering Stereo Endoscopic Attention Module (SEAM). This network aims to progressively enhance the quality of super-resolution (SR), moving from coarse to fine details. Specifically, we propose an edge-guided stereo attention mechanism integrated into each interaction of stereo features. This mechanism aims to capture consistent structural details across different views more effectively. Our proposed model demonstrates superior super-resolution reconstruction performance through comprehensive quantitative evaluations and experiments conducted on three datasets. Our E-SEVSR framework demonstrates superiority over alternative approaches. This framework leverages the edge-guided stereo attention mechanism within the multi-stage network, improving super-resolution quality in medical imaging applications.

INDEX TERMS Anatomy, stereo endoscopic attention module, stereo imaging, stereo video super-resolution.

I. INTRODUCTION

The continuous evolution of digital videos and images has led to significant advancements in visual quality across various domains. Cameras have become omnipresent, serving diverse purposes such as surveillance with CCTV, capturing moments through smartphones, contributing to medical sciences for more precise diagnostics and surgeries, aiding space exploration with satellites, and enriching daily life through various imaging devices. Technological progress has transformed imaging from the era of black and white to the current era of 8k resolution and beyond. Video and image resolution, determined by the number of pixels, is a crucial determinant of image quality.

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

Despite the impressive strides made in imaging devices and standards, certain limitations persist, resulting in blurriness of video noise and loss of vital details. In contrast to stereo images, stereo videos face heightened susceptibility to constrained spatial resolution owing to the added temporal dimension, posing potential limitations for applications demanding finer details. Despite the extensive exploration of super-resolution techniques over the years, prevailing methods predominantly center on restoring stereo images. The realm of stereo video super-resolution (StereoVSR) remains relatively uncharted. Super-resolution (SR) emerges as a solution to this challenge, involving recovering missing information from low-quality images.

Endoscopy is extensively used for surgical navigation and minimally invasive procedures [1]. However, the limited depth information and field-of-view in endoscopic videos

captured by a single camera has prompted the increasing adoption of stereo cameras, particularly in intricate and robot-assisted surgeries [2]. Stereo endoscopic images, derived from two distinct viewpoints, offer valuable depth cues and enhanced sub-pixel information compared to their single-camera counterparts [3].

However, challenges arise in endoscopy regarding maintaining high video quality and resolution, primarily due to the constraints imposed by the confined surgical space and the limited field of view of endoscopic instruments. Optical sensors must be compact to capture various tubular cavities and lumens scales effectively. Furthermore, the limitations of unstable illumination conditions can lead to image degradation and the loss of crucial information in stereo endoscopic images. These issues can negatively impact subsequent procedures such as image classification, segmentation, and reconstruction [3], [4].

Consequently, there is a considerable advantage in improving the resolution and quality of stereo endoscopic images and video frames to mitigate these challenges and enhance the overall efficacy of endoscopic procedures.

In the context of stereo endoscopic video super-resolution (SR), including consecutive frames introduces valuable temporal consistency. Traditional video SR methods typically involve the network processing several successive images, extracting and synthesizing features to reconstruct high-resolution (HR) outputs. Nonetheless, when utilizing the conventional 2D video super-resolution approach on stereo video frames, there is a potential risk of losing the alignment or correspondence between the left and right views. This risk is further elaborated in recent studies, such as the one by [80].

Despite the enhanced visual performance achieved by Convolutional Neural Network (CNN)-based super-resolution (SR) techniques compared to traditional methods, their efficacy is hampered by constraints related to the convolution kernel size and the restricted field-of-view regions. CNN models inherently possess limitations in capturing long-range dependencies. Recently, transformer networks, which integrate self-attention mechanisms, have emerged as a promising solution for addressing various visual challenges [5], [6].

Within transformer-based methodologies, the input image and frames undergo segmentation into smaller patches, which are subsequently treated as sequential token inputs. These tokenized inputs are then used to extract image features utilizing self-attention mechanisms, considering the global relationships among these tokens. The Swin transformer [7] distinguishes itself by combining the advantages of both CNN and transformer architectures through parallel computing and applying the shifted window technique. This approach builds a hierarchical feature representation, starting with small patches and gradually combining adjacent patches in deeper transformer layers. By harnessing multi-scale feature maps, the Swin transformer model efficiently and effectively employs advanced methods for dense prediction and image reconstruction [8]. Furthermore, similar principles of feature extraction and handling complex scenarios have been effectively applied in gait recognition systems, where parameters such as clothing, angle shift, and walking style significantly impact system performance. These systems utilize advanced machine learning classifiers and feature selection methods to achieve high accuracy in real-time environments, demonstrating the adaptability of transformer models in various applications [75], [76]. This integration allows for a more comprehensive consideration of global dependencies, overcoming the limitations of purely CNN-based models.

The StereoVSR task can be effectively addressed through two distinct strategies. One approach involves deploying stereo image super-resolution (StereoSR) methods [9], [10], [11], [12], [13] to independently super-resolve low-resolution (LR) observations for each frame pair. Subsequently, these reconstructed frames are merged to form high-resolution (HR) video clips. Alternatively, the second strategy utilizes cross-time information from a single view, employing video super-resolution (VSR) methods [14], [15], [16], [17], [18], [19], [20], [21], [22], [23] for spatial reconstruction. Both approaches solely merge multi-view or temporal information independently, lacking the full utilization of view-temporal correlated information between frames.

Another approach entails integrating multi-view and temporal information by first utilizing StereoSR techniques to generate super-resolved images. Following this, video super-resolution (VSR) methods are applied to enhance the high-resolution frames further, resulting in videos with even greater resolution. In contrast to the previous two methods, this approach uses information from different views and adjacent frames within a single view. However, these three approaches, while effective, do not consider cross-time-cross-view information, thereby missing an opportunity for further performance enhancement.

II. LITERATURE REVIEW

This section provides a concise overview of the super-resolution (SR) methods pertinent to our research, encompassing single-image SR [24], stereo-image SR [25], video SR techniques [26], and Stereo Video Reconstruction. While single-image super-resolution (SISR) is effective for enhancing individual images, it falls short in leveraging the continuity present in video data, such as endoscopic videos, leading to suboptimal super-resolution results. Multiple Image Super-Resolution (MISR) addresses this limitation by incorporating multiple low-resolution (LR) images to generate a single high-resolution (HR) image.

These methods typically involve taking pairs of LR and HR patches and learning mapping to translate LR patches into HR ones. Example pair techniques can be tailored for general images or specific types, such as medical images, depending on the provided training set of examples. Sparse coding stands as a state-of-the-art representative example-based super-resolution model.

In response to these challenges, researchers have introduced novel methods utilizing deep convolutional neural

networks for super-resolution reconstruction. Hu et al. [27] modified the network architecture to simplify performance and training. Another method enhances the correlation between neighboring feature information and overall image quality by incorporating context information into the network. Despite breakthroughs, ongoing research is essential to enhance super-resolution reconstruction methods further.

A. SINGLE IMAGE SRS

Single Image Super-Resolution (SISR) has been a pivotal research focus for decades, with recent advancements showcasing the efficacy of deep learning in achieving high reconstruction accuracy [28], [29], [30]. The SENext [31] approach introduces a Squeeze-and-Excitation Next architecture for SISR, leveraging squeeze-and-excitation blocks (SEB) to reduce computational costs and dynamically recalibrate channel-wise feature mappings. Utilizing local, sub-local, and global skip connections enhances feature reusability and stabilizes training convergence. SENext, employing post-upsampling in the pre-processing step, outperforms previous methods.

Kim et al. [32] propose a popular SISR technique, a Very Deep Super-Resolution Network (VDSR) with twenty layers, showcasing the increasing complexity of SR networks in exploiting intra-view information. Zhang et al. [33] fuse residual and dense connections, introducing the Residual Dense Network (RDN) for comprehensive hierarchical feature characterization. Recent advancements include Residual Channel Attention Networks (RCAN) [34], Residual Non-Local Attention Networks (RNAN) [35], and Second-order Attention Networks (SAN) [36]. Muhammad et al. [37] present a novel architecture inspired by ResNet and Xception networks, significantly reducing network parameters and enhancing processing speed while achieving high-quality HR images. Experimental results establish this technique as a state-of-the-art SR method regarding accuracy, speed, and visual quality.

B. STEREO IMAGE SR

Recently, stereo image super-resolution (SR) has gained heightened attention, with notable works exploring the effective utilization of stereo information. Enhancing stereo images requires addressing the critical challenge of efficiently applying corresponding information between two views within the SR network. Bhavsar and Rajagopalan [25] introduced a comprehensive framework designed to simultaneously estimate the image depth map and the super-resolved (SR) image using multiple low-resolution (LR) images. This framework was developed by formulating a unified energy function and iteratively minimizing it through updates to the SR image and the disparity map.

Several convolutional neural networks (CNN)–based stereo super-resolution (SR) approaches integrate features such as disparity and parallax attention. Jeon et al. [11] presented the Stereo Enhancement Super-Resolution model (StereoSR), which utilizes a single image along with a set of auxiliary shifted images to produce super-resolution (SR) results with improved details. Nonetheless, this method encountered constraints when dealing with stereo images containing varying disparities, mainly because it relied on a fixed maximum parallax. Addressing this, Wang et al. [38] introduced the Parallax-Attention Stereo Super-Resolution Network (PASSRnet). Their innovation was the introduction of the parallax attention module (PAM), which efficiently captures information from both views along the epipolar line to improve correspondence matching. Ying further expanded on this concept by integrating multiple PAMs into different stages of pre-trained single-image super-resolution (SISR) networks to enhance overall performance.

Yan et al. [9] pioneered a domain adaptive stereo super-resolution (SR) network that estimates disparities using a pre-trained stereo matching network. They harnessed cross-view information by warping views to the other side, enhancing the overall stereo SR performance. On a different note, Xu et al. [39] introduced bilateral grid processing into convolutional neural networks (CNNs), presenting a Bilateral Stereo Super-Resolution Network (BSSRnet) explicitly designed for stereo image SR. In contrast, Chu et al. [40] introduced an innovative CNN-based approach known as NAFNet, which includes a distinctive Stereo Cross Attention Module (SCAM) block designed for parallax fusion.

C. VIDEO SR

Video Super-Resolution (VSR) focuses on the task of reconstructing a high-resolution (HR) video from its corresponding low-resolution (LR) input, distinguishing itself from Single Image Super-Resolution (SISR) methods, which deal with single images [34], [41], [42], [43], Video Super-Resolution (VSR) takes advantage of temporal correlations between frames, resulting in enhanced reconstruction outcomes. In recent years, significant endeavors have been directed towards harnessing multi-frame data for VSR within deep learning [15], [20], [23], [44], [45].

A method of Video Super-Resolution (VSR) methods pertain to the alignment of various frames using motion compensation modules that rely on optical flow estimation [46], [47]. Nonetheless, the process of optical flow estimation is inherently challenging and susceptible to inaccuracies [48], [49]. Other approaches seek to exploit multiframe information implicitly [17], [23]. For instance, Wang et al. [20] employ a combination of techniques, including a pyramid, cascading, and deformable module for alignment, along with a temporal and spatial attention module for information fusion, resulting in the attainment of state-of-the-art results.

However, when dealing with the StereoVSR task, which involves input with more frames captured from an additional viewpoint, it is apparent that directly applying conventional VSR methods may not yield optimal results. To achieve improved performance, it becomes imperative to carefully

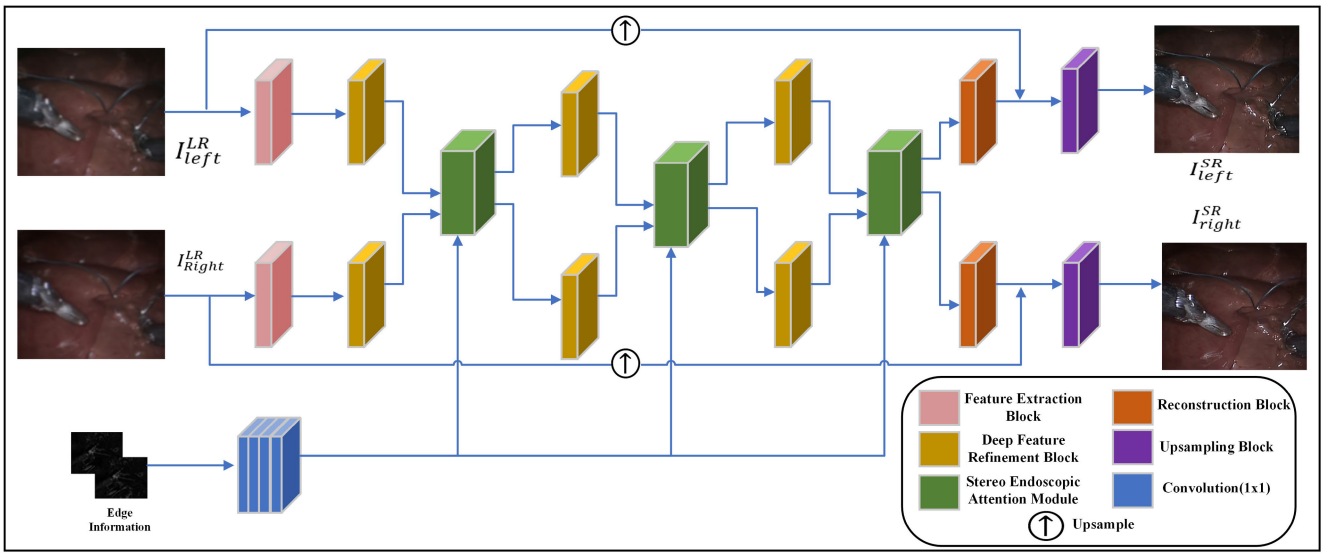

FIGURE 1. The proposed E-SEVSR network architecture provides an overview, depicted with I^{LR} representing the input low-resolution video frames on the left side, while I^{SR} represents the output reconstructed video frame on the right side. This network aims to take in low-resolution video frames and generate high-quality, super-resolved video frames as output.

account for and leverage the task-specific view-temporal correlations.

D. STEREO VIDEO SUPER-RESOLUTION

Stereo videos offer valuable multi-view information, presenting opportunities for improved performance in various reconstruction tasks. Recent studies have tackled the challenge of stereo video deblurring by simultaneously estimating 3D scene flow and removing blurs [50], [51]. Li et al. [52] have harnessed depth information for stereo video re-targeting in a different application, facilitating the seamless adjustment of stereo content to screens of varying sizes and aspect ratios.

In endoscopic surgery, the quest for enhanced visual clarity has led to developing and comparing various super-resolution (SR) methodologies, each presenting its unique strengths and weaknesses. As delineated in Table 1, beginning with the Minimally Invasive Surgery SR in 2011, which offered a comprehensive analysis of SR techniques but required a delicate balance between enhancement and artifact reduction [81]. By 2013, the RGB Hybrid 3-D Endoscopy method enhanced spatial resolution using RGB data, albeit its effectiveness was challenged in diverse environments [69]. Advancements continued with Hybrid Range Imaging and hybrid imaging with Maximum a Posteriori (MAP) estimation was introduced, though it faced motion estimation and data fusion challenges [69]. The introduction of Disparity-Constrained Parallel Attention marked a significant improvement in stereo image quality despite issues with stereo camera inconsistencies [83]. Real-Time Surgery Enhancement promised real-time performance in surgery, necessitating further data collection and testing for validation [84]. The subsequent years saw the introduction of Disparity-Constrained Stereo SR, which was effective on specific datasets but struggled with adaptability in surgical applications [63]. The most recent advancements, Channel and Spatial Attention SR and Hybrid Attention for Endoscopic Video SR highlighted the importance of detail enhancement and image reconstruction through attention mechanisms. Yet, both methodologies underscored the need for refinement in diverse settings and optimization of attention mechanisms, respectively [62], [66]. This progression underscores a continuous effort to refine image quality in endoscopic surgery, navigating the trade-offs

TABLE 1. Comparison of Super-Resolution Methods in Endoscopic Surgery.

Method	Year	Strength	Weakness
Minimally Invasive Surgery SR [81]	2011	Comprehensive analysis of SR techniques.	Balance between enhancement and artifact reduction needed.
RGB Hybrid 3-D Endoscopy [82]	2013	Enhanced spatial resolution using RGB data.	Difficult in diverse environments.
Hybrid Range Imaging [69]	2015	Hybrid imaging with MAP estimation.	Motion estimation and data fusion issues.
Disparity-Constrained Parallel Attention [83]	2020	Improved stereo image quality.	Stereo camera inconsistencies.
Real-Time Surgery Enhancement [84]	2021	Real-time performance in surgery.	Needs more data collection and testing.
Disparity-Constrained Stereo SR [63]	2022	Effective on specific datasets.	Adaptability in surgery.
Channel and Spatial Attention SR [66]	2023	Detail enhancement with attention mechanisms.	Refinement for diverse settings needed.
Hybrid Attention for Endoscopic Video SR [62]	2023	Hybrid attention for image reconstruction.	Attention mechanism optimization needed.

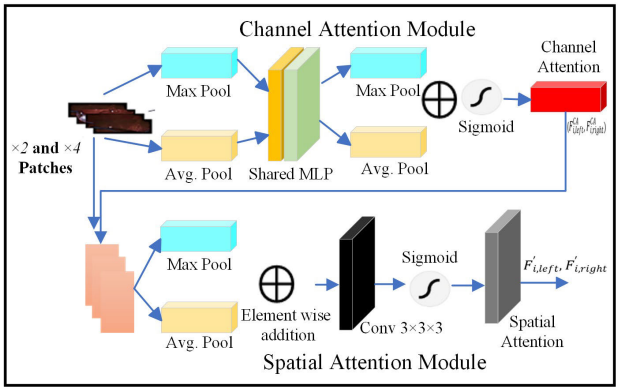

FIGURE 2. Architecture of Combined Channel and Spatial Attention Block (CCSB).

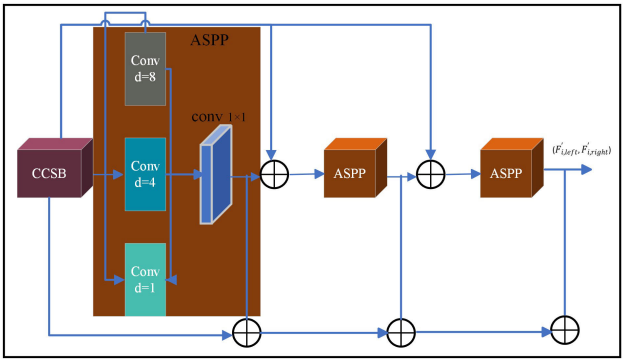

FIGURE 3. Feature Extraction block comprises one CCSB block and three ASPP blocks. ASPP has three dilated convolution layers, each with dilation rates of 1,4 and 8. It provides an enhanced receptive field to features extracted by CCSB.

between real-time applicability, environmental adaptability, and the computational demand of advanced SR techniques.

Our solution addresses the challenges previously proposed models face in the Stereo Video Super-Resolution (StereoVSR) field. This area has not been extensively explored in existing literature. This approach aims to leverage the benefits of deep learning for stereo videos, intending to achieve notably enhanced super-resolution outcomes.

III. PROPOSED METHODOLOGY

Figure. 1 illustrates our proposed Edge guided Stereo Endoscopic Video Super-Resolution (E-SEVSR) network architecture, comprising the following components: Feature Extractor (FE), Deep Feature Refinement Block (RFDB), Stereo Endoscopic Attention Module (SEAM), Reconstruction Block, and Upsampling Block.

The model $\mathcal{F}(.,.,\theta)$, governed by model parameters, is responsible for generating the reconstructed SR results $I_k^{(left,SR)}$ and $I_k^{(right,SR)}$ from the given left LR frames $\left(I_{(k-m)}^{(left,LR)},\ldots,I_k^{(left,LR)},\ldots,I_{(k+m)}^{(left,LR)}\right)$ and right LR frames $\left(I_{(k-m)}^{(right,LR)},\ldots,I_k^{(right,LR)},\ldots,I_{(k+m)}^{(right,LR)}\right)$ as represented by Equation (1):

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

A. FEATURE EXTRACTION BLOCK

In this context, superscripts indicate tensor attributes such as left (L) and right (R) views, low-resolution (LR) and super-resolution (SR) resolutions, and processing statuses by specific modules. Subscripts of tensors represent temporal information, specifically the frame count, while subscripts of modules denote their order in the process. This equation demonstrates how the model processes the input left and right LR frames to produce the corresponding high-resolution stereo endoscopic images.

The Combined Channel and Spatial Attention Block (CCSB) [66] consists of two integral components: the Channel Attention Block (CAB) and the Spatial Channel Attention Block (SAB), as depicted in Figure. 2. The CAB plays a role in determining the importance of various feature maps, while the SAB identifies critical areas within each feature map. These operations involve simultaneous average pooling and max pooling, which combine and condense the features, yielding both max-pooled and average-pooled features. During the training phase, the max-pooled and average-pooled features undergo further processing via two densely connected layers. A reduction parameter is introduced to manage parameter complexity, setting the activation size as $\frac{n_{\text{channels}}}{r\times1\times1\times1}$. Finally, a sigmoid activation function is applied, yielding channel attention values $F_i^{(left,CA)}$, $F_i^{(right,CA)}$ represents the current frame under processing.

$$F_i^{(left,CA)} = f_{\text{channel\_attention}}(I_{i-1}^{(left)}) \quad (2)$$

$$F_i^{(right,CA)} = f_{\text{channel\_attention}}(I_{i-1}^{(right)}) \quad (3)$$

Spatial attention mechanisms are explored to emphasize important regions within feature maps. The refined features obtained from channel attention are separately subjected to max pooling and global average pooling, generating a 3-dimensional feature map. Concatenating the outputs of both pooling operations, a 3-dimensional convolutional operation with a kernel size of 3×3×3 is employed, creating a three-dimensional spatial attention map. This map undergoes a sigmoid activation to yield optimized features $F'_{i,left}$, $F'_{i,right}$.

$$F'_{i,left} = f_{\text{spatial\_attention}}(F^{CA}_{i-1,left}) \quad (4)$$

$$F'_{i,right} = f_{\text{spatial\_attention}}(F^{CA}_{i-1,right}) \quad (5)$$

The low-resolution (LR) images traverse through the CAB, followed by the extracted features from the CAB passing through the SAB for further refinement.

B. DEEP FEATURE REFINEMENT BLOCK

Deep Feature Refinement Blocks (DFRB) comprise four RDB blocks each, further refining features extracted from the

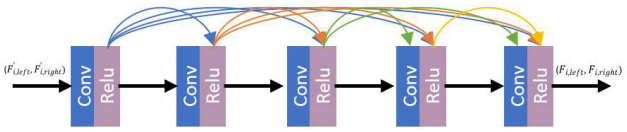

FIGURE 4. Residual Dense Blocks (RDB), consist of multiple convolutions and ReLU layers to provide deep feature extraction for feature refinement.

FE block. Utilizing Residual Dense Blocks (RDB) allows for generating numerous local features while maintaining a broad receptive field, contributing significantly to superior Super-Resolution (SR) results. Incorporating RDBs is warranted due to their inherent capacity to facilitate learning complex and hierarchical features. These blocks encompass multiple densely connected convolutional layers, fostering feature reuse and empowering the model to capture intricate patterns and structures within the data.

$$F_{i,left} = f_{\text{DFRN}}(F'_{i-1,left}) \tag{6}$$

$$F_{i,right} = f_{\text{DFRN}}(F'_{i-1,right}) \tag{7}$$

RDBs are strategically placed after the feature extraction block and after every SEAM block. Subsequently, the extracted features from these RDBs are concatenated and forwarded to the SEAM for further processing. This approach allows for comprehensive feature refinement and interaction before being utilized in the subsequent stages of the model.

C. SPATIAL FEATURE TRANSFORM

The spatial feature transform block [53] serves as a critical component in the processing pipeline, handling two sets of features: one obtained from an edge detection algorithm and the other refined through the Deep Feature Refinement Blocks (DFRB).

This block is designed to harmonize and integrate these distinct sets of features, leveraging their respective strengths. The features derived from the edge detection algorithm focus on capturing high-frequency information related to edges and boundaries within the input data. On the other hand, the refined features from the DFRB encapsulate more abstract and learned representations of the input, potentially encoding complex structures and patterns. The spatial feature transform block orchestrates the fusion or combination of these feature sets. It might employ various mechanisms, such as attention mechanisms, learnable transformations, or adaptive pooling strategies, to effectively merge the edge-focused information with the enriched and refined features from the DFRB. This fusion aims to leverage the complementary nature of the edge-derived details and the hierarchical representations learned by the DFRB.

$$F^{(m,M)}_{i,left} = (\alpha_m + 1) \odot (F_{i-1,left} + \beta_m) \tag{8}$$

$$F^{(m,M)}_{i,right} = (\alpha_m + 1) \odot (F_{i-1,right} + \beta_m) \tag{9}$$

By combining these features intelligently and synergistically, the spatial feature transform block aims to create a unified representation that encapsulates detailed edge information and abstract contextual knowledge. This consolidated feature representation can significantly enhance subsequent processing stages, contributing to the overall effectiveness and robustness of the model for the given task.

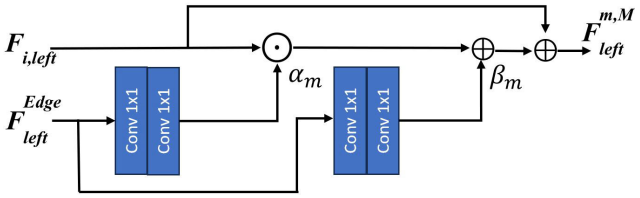

FIGURE 5. Spatial Feature Transform (SFT) block to process features from pre-trained edge detection model and output features from DFRB. SFT is deployed before every SEAM block to excel model efficiency by integrating edge information.

D. STEREO ENDOSCOPIC ATTENTION MODULE

The Super-Resolution Edge-preserving and Attention Mechanism (SEAM) system, a significant advancement in endoscopic imaging, incorporates several innovative features that set it apart from traditional super-resolution techniques. SEAM leverages stereo vision integration from stereo endoscopes, offering dual perspectives for more accurate depth and spatial relationship reconstruction. Its attention mechanism efficiently focuses on image areas with intricate details, thus enhancing super-resolution effectiveness. Unlike conventional methods, SEAM preserves edges and textures crucial for medical diagnostics and utilizes depth information from stereo images to guide the super-resolution process. This depth-aware approach is key in better preserving fine details.

Moreover, SEAM likely includes specialized noise reduction and artifact suppression components, ensuring enhanced image quality. Empirical evidence from practical applications demonstrates SEAM's superiority in preserving fine details in endoscopic images, backed by quantitative metrics and qualitative assessments.

We have also integrated the Occlusion Handling block within the SEAM module, which plays a pivotal role in generating symmetric stereo correspondence and deriving occlusions by utilizing the attention maps $M_{R\to L}$ and $M_{L\to R}$. This block is instrumental in identifying and managing occluded areas, which is crucial for accurate depth perception and feature extraction in stereo endoscopic videos.

The Occlusion Handling block's techniques are adept at processing occlusions, often caused by bodily fluids or tissues, common in endoscopic videos. These techniques focus on isolating and minimizing the impact of occlusions, leading to clearer, more interpretable images. This enhancement is vital in medical diagnostics and procedures where detail is paramount. The integration of this block improves the quality of stereo endoscopic videos by adeptly handling occlusions. It ensures accurate capture and representation of depth and spatial information, bolstering the overall effectiveness of the SEAM module in stereo endoscopic video super-resolution tasks.

Given a pair of stereo images I_L and $I_R \in \mathbb{R}^{H\times W}$, parallax attention maps $M_{R\to L}, M_{L\to R} \in \mathbb{R}^{H\times W\times W}$ can be generated. These maps are instrumental in identifying occluded regions, as they highlight areas where depth values change abruptly or near image boundaries. The occluded regions correspond to empty intervals in the attention maps, indicating the absence of counterparts in the other view.

The conversion of the right image into the left perspective, denoted as $I_{R\to L}$, is achieved through the equation:

$$I_{R\to L}(h,:) = M_{R\to L}(h,:,:) \otimes I_R(h,:) \tag{10}$$

where $\otimes$ represents batch-wise matrix multiplication. The softmax normalization performed along the third dimension of $M_{R\to L}$ and $M_{L\to R}$ indicates the matching possibility between corresponding points in the stereo images.

The possibility of a point being occluded in the right view and its effect on the left image is calculated as follows:

$$P_L(h,w_1) = \sum_{w_2=1}^{W} M_{R\to L}(h,w_1,w_2)\cdot M_{L\to R}(h,w_2,w_1) \tag{11}$$

To account for noise and rectification errors, we extend this equation by ± 2 pixels:

$$P'_L(h,w_1) = \sum_{\delta=-2}^{2}\sum_{w_2=1}^{W} M_{R\to L}(h,w_1+\delta,w_2) \cdot M_{L\to R}(h,w_2,w_1) \tag{12}$$

The valid masks V_L and V_R for the left and right views, respectively, are calculated using a tanh function applied to P'_L and P'_R, with a scaling factor τ set empirically:

$$V_L = \tanh(\tau P'_L), \quad V_R = \tanh(\tau P'_R) \tag{13}$$

The SEAM module, depicted in Figure 6, is enhanced by integrating patch-wise (PConv) and depth-wise convolution (DWConv) within stereo endoscopic attention modules for cross-view feature extraction. This integration and the Occlusion Handling block significantly advance stereo endoscopic video super-resolution image processing. It boosts the ability to streamline network architecture by reducing parameters and computational demands.

Q denotes the query matrix derived from the source intra-view feature (for example, the left-view), while K and V represent the key and value matrices derived from the target intra-view feature (for example, the right-view). The dimensions H, W, and C correspond to the feature map's height, width, and number of channels.

$$\text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{C}}\right)\cdot V \tag{14}$$

SEAM introduces a cross-view attention mechanism that amalgamates information from both left and right-view images to produce cross-view attention maps.

This strategy leverages distinct information in each view, enhancing feature fusion and improving restoration outcomes. Specifically, given the input stereo intra-view features $F_i^{(left)}$, $F_i^{(right)}$ (both with dimensions RH×W×C), the cross-view fusion features $F_{\text{left}\to\text{right}}$ are obtained through a process involving point-wise and depth-wise convolutions, denoted as $W()p$ and $W()d$, respectively. These convolutions refine features from both channel and spatial perspectives.

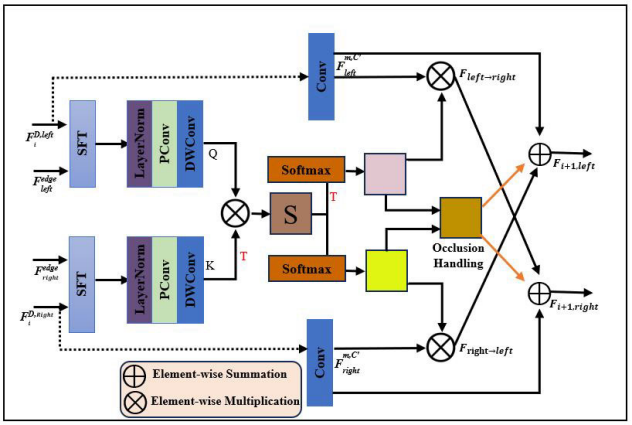

FIGURE 6. SEAM Architecture for cross-view feature extraction.

$$Q_{\text{i,left}} = W_d^{(Q_{\text{left}})} W_p^{(Q_{\text{left}})}\left(\text{LN}(F_{i,left})\right) \tag{15}$$

$$K_{\text{i,right}} = W_d^{(k_{\text{right}})} W_p^{(k_{\text{right}})}\left(\text{LN}(F_{i,right})\right) \tag{16}$$

$$V_{\text{i,right}} = W_d^{(V_{\text{right}})} W_p^{(V_{\text{right}})}(F_{i,right}) \tag{17}$$

$$F_{\text{left}\to\text{right}} = W_d^{\text{right}}\text{Attention}_{\text{left}\to\text{right}}(Q_{\text{i,left}}, K_{\text{i,right}}, V_{\text{i,right}}) \tag{18}$$

Similarly, the cross-view fusion features $F_{\text{right}\to\text{left}}$ are derived through a comparable process. Subsequently, the interacted cross-view information $F_{\text{left}\to\text{right}}$,$F_{\text{right}\to\text{left}}$ and intra-view information $F_{i,left}$, $F_{i,right}$ are fused via element-wise addition, utilizing trainable channel-wise scales denoted as γ_{left} and γ_{right}, which are initialized with zeros to stabilize training.

$$Q_{\text{i,right}} = W_d^{(Q_{\text{right}})} W_p^{(Q_{\text{right}})}\left(\text{LN}(F_{i,right})\right) \tag{19}$$

$$K_{\text{i,left}} = W_d^{(k_{\text{left}})} W_p^{(k_{\text{left}})}\left(\text{LN}(F_{i,left})\right) \tag{20}$$

$$V_{\text{i,left}} = W_d^{(V_{\text{left}})} W_p^{(V_{\text{left}})}(F_{i,left}) \tag{21}$$

$$F_{\text{right}\to\text{left}} = W_d^{\text{left}}\text{Attention}_{\text{right}\to\text{left}}(Q_{\text{i,right}}, K_{\text{i,left}}, V_{\text{i,left}}) \tag{22}$$

The final fusion equation combines these features to create a more comprehensive representation.

$$F_{i+1,left} = \gamma_{\text{left}} F_{i,(\text{left}\to\text{right})} + F_{i,left} \tag{23}$$

$$F_{i+1,right} = \gamma_{\text{right}} F_{i,(\text{right}\to\text{left})} + F_{i,right} \tag{24}$$

In summary, SEAM employs a sophisticated attention mechanism that integrates information from multiple views, enhancing feature fusion and leading to more effective restoration results. This is achieved through operations involving projections, convolutions, and fusion techniques applied to intra- and cross-view features.

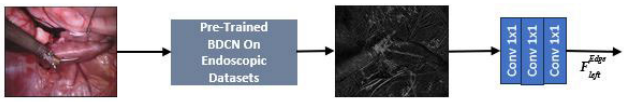

FIGURE 7. Edge Estimation using the fine-tuned BDCN model for enhanced edge detection in endoscopic imagery.

E. EDGE ESTIMATION MODULE

Drawing inspiration from [54], We have refined the BDCN [55] model, specifically tailoring it for endoscopic video super-resolution applications by fine-tuning it on various endoscopic datasets Kvasir [85], Hamlyn [86], diVinci [63], SCARED [57], and EndoVis [87]. This refinement has significantly improved the model's efficiency in detecting critical features in endoscopic imagery. Our edge estimation process involves passing the stereo input I_{LR} through the BDCN-based edge detection network [56], generating multi-scale edge probability maps (specifically at a scale of 5) for both the left and right views. These maps, now refined with the fine-tuned BDCN model, preserve stereo consistency across the views, an essential aspect of our methodology.

Subsequently, we employ a conditional subnetwork tailored for processing these enhanced edge probability maps. This subnetwork, consisting of four convolutional layers, takes as input the refined edge probability maps from both views and generates edge-guided features denoted as $F_i^{Edge} = \{F_{i,left}^{\text{Edge}}, F_{i,right}^{\text{Edge}}\}$. These features, benefiting from the improved edge detection, serve as a shared input for the cross-view interaction component.

To contain the receptive field of the conditional network and focus on the improved edge features, we opt for 1×1 kernels across all convolutional layers. This design choice minimizes interference from smooth regions within the edge probability maps, emphasizing the extraction of pertinent information associated with the enhanced edge regions. The network, thus, more effectively emphasizes edge features by employing these specific kernel sizes, allowing for a refined and selective extraction of edge-guided features critical for subsequent processing stages.

F. RECONSTRUCTION AND UPSCALING

The Reconstruction Block is the ultimate stage in the image processing pipeline, dedicated to reconstructing high-resolution (HR) images from the refined features derived from earlier processing stages. This block is meticulously crafted to enhance image quality and detail, particularly in cross-view integration for stereo endoscopic videos. Comprised of a sequence of tailored operations, the Reconstruction Block initiates with a 1×1 convolution layer ($\text{Conv}_{1\times1}$) designed to adjust channel dimensions efficiently. Following this, a Residual Dense Block (RDB) captures intricate patterns and fine-level details within the image content. The RDB's densely connected convolutional layers foster feature reuse and the extraction of intricate, hierarchical features. Subsequently, the Combined Channel and Spatial Attention

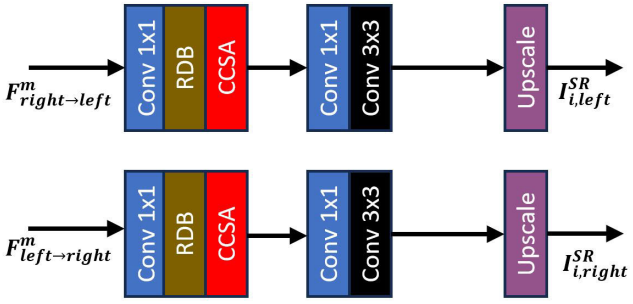

FIGURE 8. Reconstruction and Upscaling Block.

(CCSA) block is employed, enhancing the discriminative power of the reconstructed images by dynamically accentuating pertinent spatial and channel features. The output represented as $F_{i,\text{left}}^m$ for the left view's m-th block of SEAM and $F_{i,\text{right}}^m$ for the right view's m-th block of SEAM, is processed by the Reconstruction Block by feeding the output from the cross-view interaction block.

$$F_{i,left}^m = F_{\text{conv}_{3\times3}}\left(F_{\text{conv}_{1\times1}}\left(F_{\text{CCSA}}\left(F_{\text{RDB}} \times \left(F_{\text{conv}_{1\times1}}\left(\left[F_{i,left}^{(m,R)}, F_{i,right\to left}^m\right]\right)\right)\right)\right)\right) \quad (25)$$

$$F_{i,right}^m = F_{\text{conv}_{3\times3}}\left(F_{\text{conv}_{1\times1}}\left(F_{\text{CCSA}}\left(F_{\text{RDB}} \times \left(F_{\text{conv}_{1\times1}}\left(\left[F_{i,right}^{(m,R)}, F_{i,left\to right}^m\right]\right)\right)\right)\right)\right) \quad (26)$$

The CCSA layer enhances the model's capability to focus on both channel-wise and spatially relevant features, which is crucial in endoscopic video super-resolution. This dual attention mechanism aids in the recovery of intricate details and textures, which is vital for medical diagnostics. It ensures the model does not overlook subtle yet diagnostically significant details often in medical imagery.

After the CCSA layer, another $\text{Conv}_{1\times1}$ operation fine-tunes the feature representations. To further refine spatial information and ensure contextual coherence, a 3×3 convolution ($\text{Conv}_{3\times3}$) is applied. This step contributes to smoothing and enhancing local patterns, ultimately augmenting the overall quality of the super-resolved (SR) images.

$$I_{i,\text{left}}^{\text{SR}} = \text{UPSCALE}(F_{i,\text{left}}^m) \quad (27)$$

$$I_{i,\text{right}}^{\text{SR}} = \text{UPSCALE}(F_{i,\text{right}}^m) \quad (28)$$

The concluding step involves employing an Upsampling Block, which is crucial for upscaling the refined feature maps to the desired HR image size. This crucial stage reinstates LR feature maps into the HR image domain, ensuring that the final output images possess the desired level of detail and clarity.

The output of the Reconstruction Block encompasses both left and right-view SR images, representing the culmination of the entire processing pipeline. Their quality is a testament to the efficacy of the model's feature extraction, attention, and reconstruction mechanisms. Augmented with the CCSA block, the Reconstruction Block is pivotal in transforming LR stereo endoscopic inputs into high-quality, super-resolved output images.

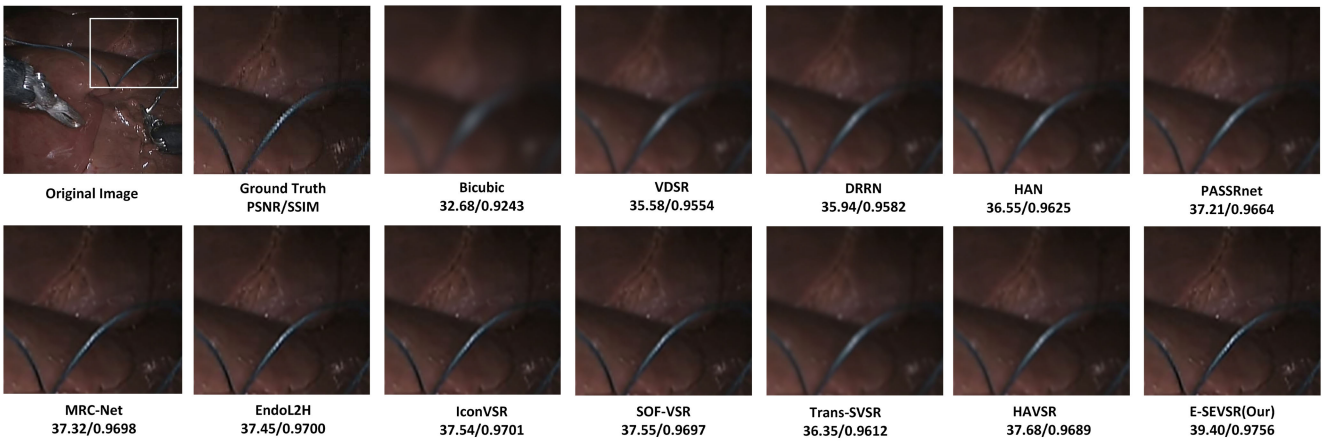

FIGURE 9. Evaluation of the perceptual quality of high-resolution images generated by image super-resolution methods for a scale factor of ×4 on di Vinci dataset.

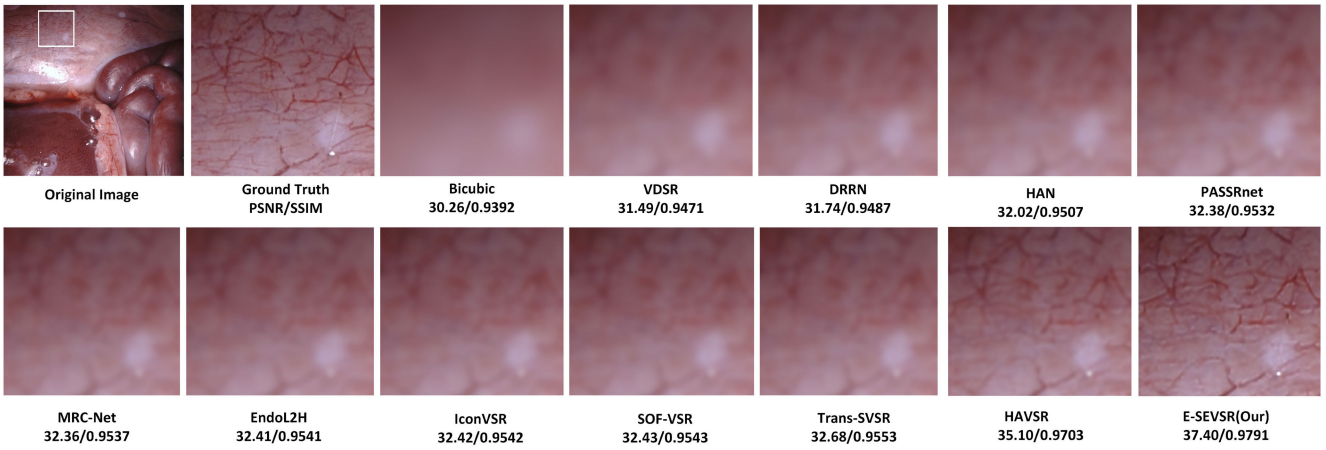

FIGURE 10. Evaluation of the perceptual quality of high-resolution images generated by image super-resolution methods for a scale factor of ×4 on the SCARED dataset.

IV. EXPERIMENTAL RESULTS

This section begins by presenting the datasets used and outlining the experimental settings. A comparative analysis is performed between the proposed model and various image SR and video SR methods. Finally, ablation studies are carried out to confirm and validate our proposed method's components and aspects.

A. EXPERIMENTAL SETTINGS

To train our model, we utilized 240 pairs of stereo video frames sourced from the da Vinci dataset [63] as the training dataset. The high-resolution (HR) images were downscaled to create low-resolution (LR) images for training using bicubic operations. Data augmentation included vertical flipping of the images. For testing, two sets of stereo endoscopic video datasets were used: the test set from the da Vinci dataset, comprising 80 pairs of stereo endoscopic video frames recorded using the da Vinci system's stereo cameras; the SCARED dataset [57], containing 120 stereo video frames; and the MICCAI 2017 Kidney Boundary Detection Sub-Challenge dataset; EndoVis dataset [85], which includes a variety of clinical conditions. This diverse testing regimen provides a comprehensive platform to evaluate the versatility and efficacy of the E-SEVSR model.

The network architecture was constructed using PyTorch and trained on an NVIDIA 3090ti GPU. For optimization, the Adam optimizer was employed with specific parameters: $\beta_1 = 0.9$ and $\beta_2 = 0.999$. A batch size of 8 was utilized during training, and the initial learning rate was set at 1×10^{-4}. In this scenario, the parameter 'k' was configured to equal 1, signifying that three consecutive frames were utilized as input data during the training process.

In our approach, we use a pixel-wise $L1$ loss function. When considering a training set with N denoting the number of training pairs, the loss function incorporating the updated parameters Θ can be expressed as follows:

$$\xi^{SR}(\Theta) = \frac{1}{N}\sum_{i=1}^{N}\left\| H_{\text{E-SEVSR}}\left(I_i^{LR} \mid \Phi\right) - I_i^{HR}\right\|_1 \qquad (29)$$

Here, Θ denotes the edge priors, which can be applied as a condition within the function. The term E-SEVSR $(\cdot)$ denotes the complete function of the proposed E-SEVSRNet,

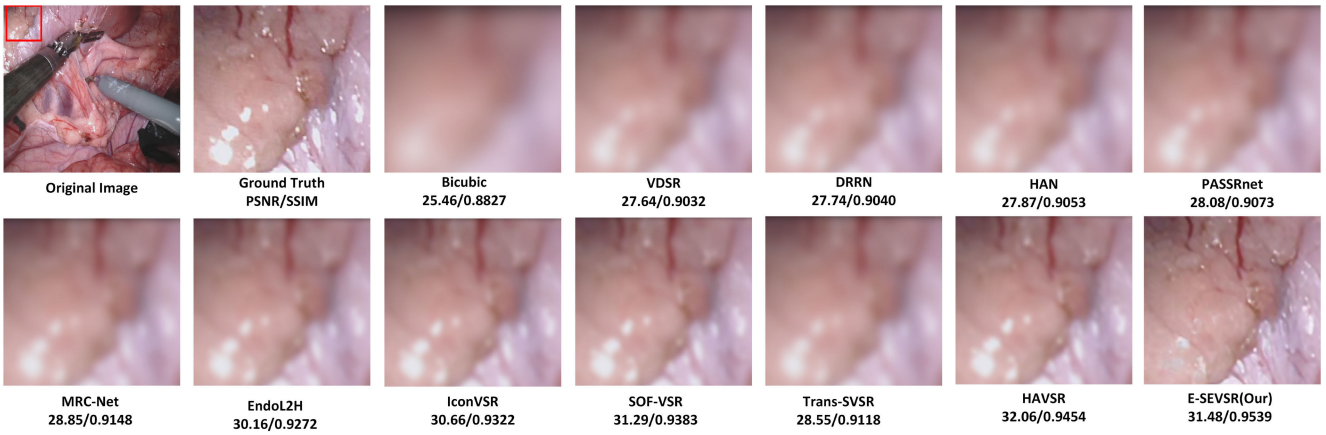

FIGURE 11. Evaluation of the perceptual quality of high-resolution images generated by image super-resolution methods for a scale factor of ×4 on the Endovis dataset.

encapsulating the entire process within the network architecture. This loss function enables the optimization of the parameters to minimize the discrepancy between the output of the E-SEVSRNet model and the ground truth high-resolution images across the training dataset.

B. EVALUATION RESULTS

Image super-resolution evaluations commonly rely on the peak signal-to-noise ratio (PSNR) as a fundamental quantitative measure to assess the similarity between high-resolution (HR) and super-resolved (SR) images. The structural similarity index measure (SSIM) is also a perceptual metric to gauge image similarity. In comparison with various SR methods, our algorithm outperformed them, with these metrics being computed within the RGB color space. The PSNR and SSIM scores are averaged across left and right image pairs among frames and calculated as (Left + Right)/2.

Table 2 presents the noteworthy PSNR and SSIM scores achieved by our proposed network on test sets for ×2 and ×4 SR tasks. Specifically, our method's PSNR values surpass those of other single, stereo, and VSR methods on the three test sets. For the ×2 stereo SR task, our model exhibits superior PSNR and SSIM values across all datasets. These quantitative evaluation results validate our model's effectiveness in leveraging temporal cross-attention and parallel attention mechanisms to reconstruct HR images.

Figure. 9, Figure. 10 and Figure. 11 depict the qualitative performance comparisons of various methods in the context of ×4 SR on the da Vinci, SCARED, and EndoVis datasets. These comparisons provide detailed observations in zoomed-in regions. Qualitatively, stereo SR methods better capture finer details than single-image super-resolution (SISR) approaches. E-SEVSR generated clearer and better-quality images than SOTA.

Figure 11 unequivocally demonstrates the efficacy of our model in environments influenced by lighting conditions on the EndoVis dataset. Notably, our model has effectively mitigated the impact of lighting variations compared to ground truth high-resolution (HR) images, showcasing its robustness in handling complex lighting scenarios. This advancement is particularly significant as lighting conditions can substantially affect super-resolved images' perceived quality and clarity.

Furthermore, compared to existing methods, our model stands out by SEAM within stereo image pairs, enhancing SR performance, especially in edge and texture details. This incorporation of SEAM contributes significantly to improving the portrayal of intricate details within the super-resolved images.

TABLE 2. Quantitative comparison using PSNR/SSIM on da Vinci dataset, SCARED and EndoVis with enlargement factor ×2 and ×4.

Method	Scale	da Vinci	SCARED	EndoVis
bicubic [58]	×2	37.66/0.9645	38.60/0.9792	31.42/0.9613
VDSR [32]	×2	37.10/0.9681	39.57/0.9824	34.92/0.9775
DRRN [59]	×2	37.98/0.9733	40.18/0.9858	35.33/0.9859
HAN [60]	×2	38.25/0.9765	40.62/0.9869	35.99/0.9887
PASSRNet [38]	×2	37.65/0.9714	40.36/0.9860	36.28/0.9897
MRC-Net [77]	×2	37.78/0.9715	40.39/0.9861	36.82/0.9893
EndoL2H [72]	×2	37.87/0.9715	40.46/0.9863	37.01/0.9898
IconVSR [78]	×2	37.99/0.9723	40.53/0.9867	37.06/0.9898
SOF-VSR [79]	×2	38.07/0.9727	40.61/0.9869	37.11/0.9899
Trans-SVSR [61]	×2	38.21/0.9767	40.73/0.9875	36.31/0.9896
HA-VSR [62]	×2	38.37/0.9771	40.80/0.9870	39.81/0.9913
E-SEVSR (Our)	×2	**43.19/0.9946**	**42.11/0.9925**	**40.43/0.9919**
bicubic [58]	×4	30.06/0.9358	32.85/0.9480	25.45/0.9161
VDSR [32]	×4	31.14/0.9410	33.35/0.9516	26.87/0.9346
DRRN [59]	×4	31.67/0.9428	34.01/0.9558	26.92/0.9353
HAN [60]	×4	31.74/0.9433	34.50/0.9569	26.99/0.9362
PASSRNet [38]	×4	31.46/0.9415	34.12/0.9547	27.10/0.9375
MRC-Net [77]	×4	31.50/0.9416	34.18/0.9548	27.48/0.9422
EndoL2H [72]	×4	31.63/0.9421	34.31/0.9554	28.01/0.9493
IconVSR [78]	×4	31.72/0.9420	34.44/0.9540	28.18/0.9519
SOF-VSR [79]	×4	31.74/0.9449	34.53/0.9559	28.38/0.9551
Trans-SVSR [61]	×4	31.89/0.9469	34.68/0.9573	27.33/0.9403
HA-VSR [62]	×4	32.03/0.9477	34.79/0.9576	28.59/0.9585
E-SEVSR (Our)	×4	**33.97/0.9721**	**36.44/0.9719**	**32.81/0.9753**

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

	Conv	CCSB	ASPP	PSNR/SSIM
Feature Extraction Block	✓	×	×	42.81/0.9941
	×	✓	×	42.91/0.9942
	×	✓	✓	**43.07/0.9943**

V. THE SIGNIFICANCE OF EDGES IN ENDOSCOPIC IMAGE AND VIDEO SUPER-RESOLUTION

Recent advancements in edge enhancement techniques have significantly improved the sharpness and clarity of endoscopic images [67]. Techniques such as edge enhancement optimization increase perceptual sharpness and reduce noise, thereby improving the overall image quality perceived by medical professionals. This is particularly important in endoscopic procedures where fine details and contrasts in tissue structures play a critical role in diagnosis [67]. Edges play a crucial role in the processing and analysis of digestive endoscopy images, the diagnosis of colorectal diseases, and the identification of pathological collagen [68]. Enhanced edge representation is vital for distinguishing between tissue types and identifying abnormalities. This becomes even more crucial in minimally invasive surgeries, where visual clarity and detail are paramount for successful outcomes [69]. Edge enhancement in endomicroscopy is crucial because it leads to more precise visualization of cellular structures and tissues [70]. In conclusion, recognizing that endoscopic videos are essentially sequences of image frames, the role of edge enhancement becomes doubly significant in both endoscopic images and video super-resolution. Edges are crucial in delineating critical structures in each frame, impacting the overall effectiveness of video analysis and diagnostics. Our model, by introducing a novel edge detection technique for endoscopic video super-resolution, has demonstrated improved results both quantitatively and qualitatively.

VI. ABLATION STUDY

A. FE BLOCK

Our model's effectiveness undergoes validation through diverse feature extraction techniques: Conv, CCSB, and CCSB+ASPP. These techniques extract features utilized for subsequent transformation. Table 3 results highlight the superior performance achieved when CCSB collaborates with ASPP, showcasing PSNR/SSIM scores of 43.07/0.9943.

Notably, omitting ASPP from the FE process significantly impacts performance. The absence of ASPP leads to a noticeable decrease in PSNR and SSIM, dropping by 0.16 dB and 0.0001, respectively. Furthermore, restricting the feature extraction to Conv, as employed in MESFINet [54], results in a more substantial decline in PSNR and SSIM, with a collective decrease of 0.26 dB and 0.0002, respectively. This notable decrease significantly impacts overall performance, emphasizing the critical role of combining CCSB and ASPP to achieve optimal outcomes.

The quantitative outcomes effectively emphasize the benefits and effectiveness of integrating CCSB and ASPP simultaneously, reaffirming their pivotal contribution to substantial performance improvements. These findings underscore the crucial nature of this feature extraction strategy within our model, highlighting its capability to enhance output quality significantly.

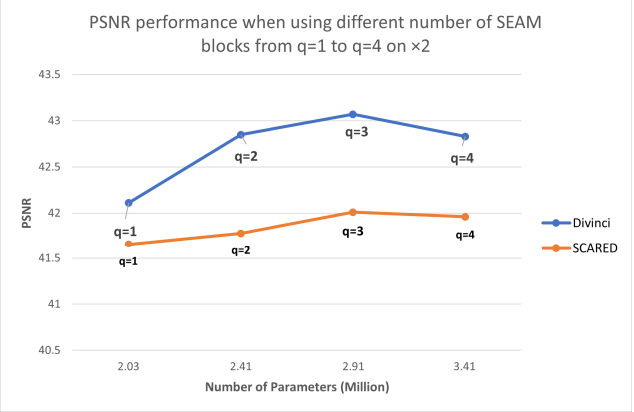

FIGURE 12. PSNR Performance when using different number of SEAM blocks from q=1 to q=4 onusing di Vinci dataset on ×2.

TABLE 4. Ablation Study by increasing the number of SEAM blocks from $q = 1$ to $q = 4$.

Number of SEAM blocks	Scale	da Vinci	SCARED
$q = 1$	×2	42.11/0.9931	41.66/0.9917
$q = 2$	×2	42.85/0.9939	41.78/0.9918
$q = 3$	×2	**43.07/0.9943**	**42.01/0.9931**
$q = 4$	×2	42.83/0.9943	41.96/0.9331
$q = 1$	×4	33.61/0.9707	35.15/0.9749
$q = 2$	×4	33.72/0.9715	36.11/0.9755
$q = 3$	×4	**33.90/0.9719**	**36.41/0.9762**
$q = 4$	×4	33.65/0.9710	35.71/0.9720

B. NUMBER OF SEAM BLOCK

We began our exploration by examining the impact of varying the number of SEAM blocks within the network while keeping the number of RDBs fixed at 4. Figure. 11 displays the trade-off between PSNR and network parameters across different quantities of SEAMs. The results, centered on both at ×2 and ×4, are detailed in Table 4, where we conducted an ablation study by progressively increasing the integration of SEAM blocks into the network, varying from q=1 to q=4.

Our analysis indicates that setting q=3 strikes an optimal balance between SR performance and network parameters. This configuration allows for consistent enhancements by leveraging additional stereo information for image reconstruction. To optimize this equilibrium, we ultimately adopt a 3-stage E-SEVSR.

C. EDGE PROBABILITY MAP

In exploring the impact of edge probability maps on Image Super-Resolution (SR), we utilized diverse edge detectors to generate these maps, presenting the outcomes in Table 5. Our analysis, derived from the tabulated data, underscores the pivotal influence of high-quality edge priors in shaping SR performance.

TABLE 5. Ablation Study incorporating different edge detection models using di Vinci dataset on ×2 .

Edge Detector	PSNR	SSIM
Sobel [64]	41.43	0.9940
Canny [65]	42.55	0.9941
DexiNed [73]	42.63	0.9941
RCN [74]	42.95	0.9942
BDCN(Fine Tunned)	**43.19**	**0.9946**

Table 5 demonstrates the intrinsic connection between the quality of edge priors and the overall SR performance. Notably, an enhancement in the quality of edge probability maps correlates with superior SR outcomes. Interestingly, while differences among various detectors' edge probability maps are discernible, the impact of the detector choice appears somewhat limited.

Specifically, analyzing edge probability maps generated by Canny [65], Sobel [64], DexiNed [73], RCN [74] and BDCN (Fine Tunned), we note BDCN's significant impact on PSNR, showcasing a remarkable increase of 0.24 dB compared to RCN [74]. Additionally, a rise of 0.0004 in SSIM is evident. Consequently, our model incorporates BDCN (Fine-tuned) for edge estimation, acknowledging its crucial role in augmenting model performance. These observations underscore the pivotal significance of edge priors, with BDCN (Fine Tunned) exhibiting a notable advantage in this context.

VII. LIMITATIONS AND FUTURE WORK

Our current model establishes a robust baseline for Stereo Endoscopic Video Super-Resolution, adhering to experimental procedures paralleled in existing studies [62], [71], [72]. However, our model currently does not support real-time super-resolution in endoscopic surgeries due to computational constraints and the lack of resources for real-time application in surgical settings. Future enhancements could include integrating motion estimation blocks, frame interpolation, and feature temporal interpolation. Additionally, hardware improvements such as using multiple GPUs, high-speed I/O interfaces, FPGA, server clustering, or Application-Specific Integrated Circuits (ASIC) could significantly augment real-time processing capabilities.

While our model is currently specialized for endoscopy, adapting it for broader applications in medical imaging, including modalities such as MRI, CT, and PET, is a compelling direction for our future research. These potential modifications and advancements pave the way for the practical deployment of our model in real-time surgical environments and beyond, extending its applicability and efficacy in clinical settings.

VIII. CONCLUSION

Our paper introduces a novel Stereo Endoscopic Attention Module (SEAM) to enhance cross-view feature interaction in Video Super-Resolution (VSR). To further augment stereo SR performance, we propose integrating a pre-trained BDCN (Fine-tuned) model to leverage edge information effectively. We demonstrate the effectiveness of our proposed network by conducting comprehensive comparisons, both qualitatively and quantitatively, with existing models in the domain of stereo super-resolution. These experiments are designed to illustrate the superior performance of our model, showcasing its competitive advantage over other methodologies in the field in terms of visual quality and quantitative evaluation metrics. Moreover, we substantiate the effectiveness of our SEAM through a series of experiments that involve quantitative comparisons. These experiments highlight the advantages and improvements of incorporating our proposed Stereo Endoscopic Attention Module. This demonstrates its capability to significantly enhance the quality and performance of stereo super-resolution tasks compared to other existing methods.

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] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 15908–15919.

[6] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He, "A survey of visual transformers," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–21, 2023, doi: 10.1109/TNNLS.2022.3227717.

[7] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[8] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.

[9] 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.

[10] X. Ying, Y. Wang, L. Wang, W. Sheng, W. An, and Y. Guo, "A stereo attention module for stereo image super-resolution," *IEEE Signal Process. Lett.*, vol. 27, pp. 496–500, 2020.

[11] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1721–1730.

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

[13] 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.

[14] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi, "Real-time video super-resolution with spatio-temporal networks and motion compensation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2848–2857.

[15] D. Fuoli, S. Gu, and R. Timofte, "Efficient video super-resolution through recurrent latent space propagation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3476–3485.

[16] M. Haris, G. Shakhnarovich, and N. Ukita, "Recurrent back-projection network for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3892–3901.

[17] Y. Jo, S. W. Oh, J. Kang, and S. J. Kim, "Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3224–3232.

[18] J. Luo, S. Huang, and Y. Yuan, "Video super-resolution using multi-scale pyramid 3D convolutional networks," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1882–1890.

[19] Y. Tian, Y. Zhang, Y. Fu, and C. Xu, "TDAN: Temporally-deformable alignment network for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3357–3366.

[20] 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.

[21] Z. Xiao, X. Fu, J. Huang, Z. Cheng, and Z. Xiong, "Space-time distillation for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2113–2122.

[22] Z. Xiao, Z. Xiong, X. Fu, D. Liu, and Z.-J. Zha, "Space-time video super-resolution using temporal profiles," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 664–672.

[23] P. Yi, Z. Wang, K. Jiang, J. Jiang, and J. Ma, "Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3106–3115.

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

[25] 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.

[26] 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.

[27] H. Hu, S. Yang, X. Li, Z. Cheng, T. Liu, and J. Zhai, "Polarized image super-resolution via a deep convolutional neural network," *Opt. Exp.*, vol. 31, no. 5, p. 8535, 2023.

[28] K. Zhang, L. Van Gool, and R. Timofte, "Deep unfolding network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3214–3223.

[29] 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.

[30] 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. 4915–4924.

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

[32] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.

[33] 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.

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

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

[36] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11057–11066.

[37] 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.

[38] L. Wang, Y. Wang, Z. Liang, Z. Lin, J. Yang, W. An, and Y. Guo, "Learning parallax attention for stereo image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12242–12251.

[39] 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.

[40] X. Chu, L. Chen, and W. Yu, "NAFSSR: Stereo image super-resolution using NAFNet," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1238–1247.

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

[42] 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.

[43] 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.

[44] D. Liu, Z. Wang, Y. Fan, X. Liu, Z. Wang, S. Chang, and T. Huang, "Robust video super-resolution with learned temporal dynamics," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2526–2534.

[45] X. Ying, L. Wang, Y. Wang, W. Sheng, W. An, and Y. Guo, "Deformable 3D convolution for video super-resolution," *IEEE Signal Process. Lett.*, vol. 27, pp. 1500–1504, 2020.

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

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

[48] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1647–1655.

[49] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8934–8943.

[50] 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.

[51] 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.

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

[53] X. Wang, K. Yu, C. Dong, and C. Change Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 606–615.

[54] 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.

[55] J. He, S. Zhang, M. Yang, Y. Shan, and T. Huang, "Bi-directional cascade network for perceptual edge detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3823–3832.

[56] C.-Y. Yang and M.-H. Yang, "Fast direct super-resolution by simple functions," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 561–568.

[57] M. Allan et al., "Stereo correspondence and reconstruction of endoscopic data challenge," 2021, *arXiv:2101.01133*.

[58] J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[59] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2790–2798.

[60] M. Cheng, H. Ma, Q. Ma, X. Sun, W. Li, Z. Zhang, X. Sheng, S. Zhao, J. Li, and L. Zhang, "Hybrid transformer and CNN attention network for stereo image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 1702–1711.

[61] 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. 705–714.
[62] 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.
[63] T. Zhang, Y. Gu, X. Huang, J. Yang, and G.-Z. Yang, "Disparity-constrained stereo endoscopic image super-resolution," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 17, no. 5, pp. 867–875, May 2022.
[64] I. Sobel et al., "A 3×3 isotropic gradient operator for image processing," Stanford Artif. Project, 1968, pp. 271–272.
[65] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.
[66] M. Hayat, S. Armvith, and T. Achakulvisut, "Combined channel and spatial attention-based stereo endoscopic image super-resolution," in *Proc. TENCON*, Jun. 2023, pp. 920–925, doi: 10.1109/TENCON58879.2023.10322331.
[67] G. Geleijnse, L. L. Veder, M. M. Hakkesteegt, H. H. W. de Gier, B. Rieger, and R. M. Metselaar, "Edge enhancement optimization in flexible endoscopic images to the perception of ear, nose and throat professionals," *Laryngoscope*, vol. 134, no. 2, pp. 842–847, Feb. 2024.
[68] L. Yang, Z. Li, S. Ma, and X. Yang, "Artificial intelligence image recognition based on 5G deep learning edge algorithm of digestive endoscopy on medical construction," *Alexandria Eng. J.*, vol. 61, no. 3, pp. 1852–1863, Mar. 2022.
[69] T. Köhler, S. Haase, S. Bauer, J. Wasza, T. Kilgus, L. Maier-Hein, C. Stock, J. Hornegger, and H. Feußner, "Multi-sensor super-resolution for hybrid range imaging with application to 3-D endoscopy and open surgery," *Med. Image Anal.*, vol. 24, no. 1, pp. 220–234, 2015.
[70] C. Zhang, Y. Gu, and G.-Z. Yang, "Contrastive adversarial learning for endomicroscopy imaging super-resolution," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 8, pp. 3994–4005, 2023, doi: 10.1109/JBHI.2023.3275563.
[71] X. Song, H. Tang, C. Yang, G. Zhou, Y. Wang, X. Huang, J. Hua, G. Coatrieux, X. He, and Y. Chen, "Deformable transformer for endoscopic video super-resolution," *Biomed. Signal Process. Control*, vol. 77, Aug. 2022, Art. no. 103827.
[72] Y. Almalioglu, K. B. Ozyoruk, A. Gokce, K. Incetan, G. I. Gokceler, M. A. Simsek, K. Ararat, R. J. Chen, N. J. Durr, F. Mahmood, and M. Turan, "EndoL2H: Deep super-resolution for capsule endoscopy," *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 4297–4309, Dec. 2020.
[73] X. Soria, E. Riba, and A. Sappa, "Dense extreme inception network: Towards a robust CNN model for edge detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1912–1921.
[74] A. Peter Kelm, V. Soorya Rao, and U. Zoelzer, "Object contour and edge detection with RefineContourNet," 2019, *arXiv:1904.13353*.
[75] M. I. Sharif, M. A. Khan, A. Alqahtani, M. Nazir, S. Alsubai, A. Binbusayyis, and R. Damaševič ius, "Deep learning and kurtosis-controlled, entropy-based framework for human gait recognition using video sequences," *Electronics*, vol. 11, no. 3, p. 334, 2022.
[76] M. Imran Sharif, M. Mehmood, M. Irfan Sharif, and M. Palash Uddin, "Human gait recognition using deep learning: A comprehensive review," 2023, *arXiv:2309.10144*.
[77] Z. Chen, X. Guo, P. Y. M. Woo, and Y. Yuan, "Super-resolution enhanced medical image diagnosis with sample affinity interaction," *IEEE Trans. Med. Imag.*, vol. 40, no. 5, pp. 1377–1389, May 2021.
[78] K. C. K. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "BasicVSR: The search for essential components in video super-resolution and beyond," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4945–4954.
[79] L. Wang, Y. Guo, L. Liu, Z. Lin, X. Deng, and W. An, "Deep video super-resolution using HR optical flow estimation," *IEEE Trans. Image Process.*, vol. 29, pp. 4323–4336, 2020.
[80] Y. Akkem, S. K. Biswas, and A. Varanasi, "Smart farming using artificial intelligence: A review," *Eng. Appl. Artif. Intell.*, vol. 120, Apr. 2023, Art. no. 105899.
[81] V. De Smet, V. Namboodiri, and L. Van Gool, "Super-resolution techniques for minimally invasive surgery," in *Proc. AE-CAI*, 2011, pp. 41–50.
[82] T. Köhler, S. Haase, S. Bauer, J. Wasza, T. Kilgus, L. Maier-Hein, H. Feußner, and J. Hornegger, "ToF meets RGB: Novel multi-sensor super-resolution for hybrid 3-D endoscopy," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22–26, 2013, Proceedings, Part I 16*. Berlin, Germany: Springer, 2013, pp. 139–146.
[83] T. Zhang, Y. Gu, X. Huang, E. Tu, and J. Yang, "Stereo endoscopic image super-resolution using disparity-constrained parallel attention," 2020, *arXiv:2003.08539*.
[84] R. Wang, D. Zhang, Q. Li, X.-Y. Zhou, and B. Lo, "Real-time surgical environment enhancement for robot-assisted minimally invasive surgery based on super-resolution," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 3434–3440.
[85] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt, M. Riegler, and P. Halvorsen, "KVASIR: A multi-class image dataset for computer aided gastrointestinal disease detection," in *Proc. 8th ACM Multimedia Syst. Conf.* New York, NY, USA: ACM, Jun. 2017, pp. 164–169, doi: 10.1145/3083187.3083212.
[86] T. H. Centre. (2020). *Hamlyn Datasets*. [Online]. Available: http://hamlyn.doc.ic.ac.uk/vision/data/daVinci.zip
[87] M. Allan, A. Shvets, T. Kurmann, Z. Zhang, R. Duggal, Y.-H. Su, N. Rieke, I. Laina, N. Kalavakonda, S. Bodenstedt, L. Herrera, W. Li, V. Iglovikov, H. Luo, J. Yang, D. Stoyanov, L. Maier-Hein, S. Speidel, and M. Azizian, "2017 robotic instrument segmentation challenge," 2019, *arXiv:1902.06426*.

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's in 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. In addition, he was honored with the Best Conference Paper Award from the 2023 IEEE Region 10 Conference (TENCON), held in 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, WA, USA, in 1996 and 2001, respectively. She joined Chulalongkorn University, in June 2001. 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 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 a former Technical Committee Chair of 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.