Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.052217

ARTICLE

CMMCAN: Lightweight Feature Extraction and Matching Network for Endoscopic Images Based on Adaptive Attention

Nannan Chong[1,2,*] and Fan Yang[1]

[1]School of Electronic and Information Engineering, Hebei University of Technology, Tianjin, 300401, China

[2]School of Information and Intelligence Engineering, Tianjin Renai College, Tianjin, 301636, China

*Corresponding Author: Nannan Chong. Email: chongnannan@163.com

Received: 26 March 2024 Accepted: 24 June 2024 Published: 15 August 2024

ABSTRACT

In minimally invasive surgery, endoscopes or laparoscopes equipped with miniature cameras and tools are used to enter the human body for therapeutic purposes through small incisions or natural cavities. However, in clinical operating environments, endoscopic images often suffer from challenges such as low texture, uneven illumination, and non-rigid structures, which affect feature observation and extraction. This can severely impact surgical navigation or clinical diagnosis due to missing feature points in endoscopic images, leading to treatment and postoperative recovery issues for patients. To address these challenges, this paper introduces, for the first time, a Cross-Channel Multi-Modal Adaptive Spatial Feature Fusion (ASFF) module based on the lightweight architecture of EfficientViT. Additionally, a novel lightweight feature extraction and matching network based on attention mechanism is proposed. This network dynamically adjusts attention weights for cross-modal information from grayscale images and optical flow images through a dual-branch Siamese network. It extracts static and dynamic information features ranging from low-level to high-level, and from local to global, ensuring robust feature extraction across different widths, noise levels, and blur scenarios. Global and local matching are performed through a multi-level cascaded attention mechanism, with cross-channel attention introduced to simultaneously extract low-level and high-level features. Extensive ablation experiments and comparative studies are conducted on the HyperKvasir, EAD, M2caiSeg, CVC-ClinicDB, and UCL synthetic datasets. Experimental results demonstrate that the proposed network improves upon the baseline EfficientViT-B3 model by 75.4% in accuracy (Acc), while also enhancing runtime performance and storage efficiency. When compared with the complex DenseDescriptor feature extraction network, the difference in Acc is less than 7.22%, and IoU calculation results on specific datasets outperform complex dense models. Furthermore, this method increases the F1 score by 33.2% and accelerates runtime by 70.2%. It is noteworthy that the speed of CMMCAN surpasses that of comparative lightweight models, with feature extraction and matching performance comparable to existing complex models but with faster speed and higher cost-effectiveness.

KEYWORDS

Feature extraction and matching; lightweighted network; medical images; endoscopic; attention

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1 Introduction

Minimally invasive surgery, as a pivotal advancement in modern medicine, has found widespread application in clinical practice. By utilizing small incisions or natural body cavities for therapeutic interventions, minimally invasive surgery reduces trauma, postoperative pain, and the occurrence of complications, while also shortening patient recovery times and improving surgical safety and efficacy. Leveraging advanced surgical instruments and sensor-guided imaging technologies such as endoscopes or laparoscopes, minimally invasive surgery ensures the accuracy and success rate of procedures. However, the intraoperative visual quality in minimally invasive surgery is often influenced by factors such as hardware performance, optical lens quality, and visual processing algorithms, and is susceptible to lighting conditions, noise, and non-rigid structures of observed objects. Consequently, efficient and accurate extraction of features from endoscopic images and real-time matching have been longstanding research concerns among scholars.

Traditional feature extraction methods for medical endoscopic images mostly rely on handcrafted features (such as SIFT, SURF, ORB, etc.). These methods are sensitive to factors like lighting, viewpoint, occlusion, and often require manual parameter tuning to adapt to different scenes and datasets. Alternatively, machine learning-based approaches (such as SVM, Random Forest) suffer from the need for extensive labeled data for training, and their performance heavily depends on the choice of feature extraction and classification algorithms, limiting their flexibility. Deep learning-based methods (such as CNNs, RNNs) require substantial labeled data and computational resources for training. Additionally, these models have large parameter sizes, which may not be suitable for resource-constrained environments.

Based on neural networks, the extraction of medical image features demands feasibility, real-time capability, and high accuracy, especially in applications like clinical medicine and surgical navigation where endoscopic imaging is employed. However, the significant computational and storage requirements associated with these networks pose substantial challenges for their deployment in medical robotics. Motivated by this practical challenge, we design efficient feature extraction and fusion modules tailored for resource-constrained endoscopic images and surgical navigation systems, characterized by limitations such as weak texture, uneven illumination, and non-rigid structures. These modules aim to provide denser, more accurate, and real-time feature data for subsequent tasks. Lightweight networks such as the SqueezeNet series boast smaller model sizes by reducing parameter count using 1×1 convolutional kernels, yet they may sacrifice some accuracy in certain complex tasks. The ShuffleNet series, on the other hand, employs an efficient channel shuffling mechanism to reduce computational complexity and model size, though it may not outperform other models in certain tasks. MobileNet series reduces model size by utilizing depthwise separable convolutions to decrease computational load, albeit potentially sacrificing accuracy compared to some tasks. The IGCV series combines Inception structures to enhance model performance and versatility across various tasks, albeit with larger model sizes compared to others. EfficientViT introduces a compound scaling method, which optimally selects scaling ratios for width, depth, and resolution dimensions, enabling the model to achieve higher accuracy. Building upon these considerations, this paper proposes a method for feature extraction and matching in monocular medical endoscopic images based on adaptive attention mechanisms. The aim is to perform feature extraction and matching in real-time, effectively, and densely. The contribution of this paper lies in:

- **Lightweight Feature Extraction and Matching Network CMMCAN:** Inspired by EfficientViT lightweight convolutional neural networks, this paper proposes a cross-modal and cross-stage multi-level cascaded adaptive attention feature extraction and matching network CMMCAN

based on an encoder-decoder structure for medical endoscopic images. For the first time, this network calculates grayscale and optical flow information, adapts attention weights hierarchically between layers, and introduces context feature guidance and aggregation between encoders and decoders, integrating features extracted from multiple layers from low-level to high-level, local to global, especially suitable for handling medical endoscopic images with low texture and uneven illumination.

- **Cross-Channel Cross-Stage Adaptive Attention Module ASFF:** This paper proposes for the first time a lightweight module ASFF based on a Siamese network for cross-channel and cross-stage adaptive attention. This module extracts from low-level local information to high-level semantic features in each branch of the dual-branch network. An adaptive attention module SCIM is introduced between layers of the dual-branch network, which adaptively adjusts attention weights at different levels to extract cross-modal features of grayscale and optical flow, fully utilizing the static and dynamic information of endoscopic images, thereby making the network have photometric consistency and texture robustness, and also improving its understanding of limited endoscopic data.
- **Global Context Feature Guidance and Aggregation Module GCGFA:** Between Encoder and Decoder: By guiding the processing of the decoder and gradually integrating cross-modal and cross-channel multi-level cascaded features, cascading and feature aggregation from local to global features, from low-level to high-level features are performed. Since this module processes low to high-level features extracted from the five-level extraction, and does not start training from the original data with a large amount of redundant information of unstructured features, the training and inference time costs are not high, thereby ensuring the lightweight of the entire network.
- **Dataset:** In addition to the publicly available medical endoscopic datasets Hyperkvasir, EAD, M2caiSeg, CVC-ClinicDB, and UCL synthetic datasets, we also collected clinical endoscopic videos and CT data from 83 anonymous patients from Tianjin Academy of Traditional Chinese Medicine Affiliated Hospital, forming an experimental dataset for model training and quantitative/qualitative evaluation.

2 Related Works

2.1 Feature Extraction Related Technology

Feature extraction stands as the most critical stage in automated machine learning methods. Significant progress has been made in improving the efficiency of this stage over the past few years. Features can vary in type, including statistical features such as mean, standard deviation, skewness, kurtosis, geometric features like area, perimeter, circularity, equivalent diameter, texture features, color features, multi-resolution features, etc. Although most feature descriptors were initially developed for computer vision tasks with natural images, these descriptors have been widely utilized in medical image analysis problems such as classification and lesion detection. The most common feature extraction techniques include Local Binary Patterns [1], Oriented Gradient Histograms [2], Gray-Level Co-occurrence Matrices [3], Discrete Cosine Transform [4], Scale-Invariant Feature Transform [5], Discrete Wavelet Transform [6], Curvelets [7], etc. Features obtained using these methods are referred to as handcrafted or engineered features. However, certain features obtained using different feature extraction methods may be redundant or irrelevant for specific tasks, leading to dimensionality reduction and performance degradation. Feature selection techniques play a crucial role in addressing these issues by selecting the most suitable data representation. These techniques aid in understanding the data, reducing computation time, and avoiding the curse of dimensionality. The most commonly

used methods are filter, wrapper, and embedded methods. Filter methods use statistical metrics to select features based on their intrinsic properties and then use the selected features to train predictors. In contrast, wrapper and embedded methods optimize objective functions to find feature subsets that offer the highest predictor performance. Embedded methods perform feature selection and algorithm training in parallel. Additionally, some dimensionality reduction methods, such as Principal Component Analysis, Linear Discriminant Analysis, etc., have been widely applied in medical image analysis tasks. In medical image analysis, selecting a good feature descriptor for a specific task heavily relies on domain knowledge and remains a challenging task [8]. Therefore, feature learning-based methods have recently garnered development in this field.

2.2 Transformer with Medical Images

To enhance feature extraction efficiency and improve downstream task performance, many scholars consider introducing the self-attention mechanism of Transformers for intelligent processing of medical images. Liao et al. [9] proposed a Multiscale Context Fusion (MSCF) module, which constructs a pyramid pooling structure and an anisotropic strip pooling structure using pooling kernels of different sizes and shapes. Pyramid pooling can extract features with different receptive fields, enriching feature representation, while anisotropic strip pooling can establish long-range dependencies from different directions, enhancing the recognition ability of elongated organs. Compared to other feature extraction modules such as Transformers and expanded convolution, MSCF can establish long-range dependencies in specific directions with fewer parameters and floating-point operations, and can more effectively handle irregularly shaped organs. A Dual Self-Attention (DSA) [10] module was developed, establishing global information connections from both spatial and channel domains. This module utilizes shift convolution to extract spatial features from each channel's feature map. Shift convolution consists of multiple hot operators and does not involve any trainable parameters. Residual modules are designed in skip connections to compensate for information loss caused by downsampling and avoid redundant transmission of shallow features. The residual module enables the framework to focus more on important features such as small targets and image edges. Fan proposed a CNN-based U-Net backbone and SA parallel network ccap-unet. The encoder comprises two parallel branches of CNN and Transformer, extracting features from input images while considering global dependencies and local information. Since medical images come from specific frequency bands within the spectrum, their color channels are not as uniformly distributed as natural images. Moreover, medical segmentation focuses more on the lesion regions in images. The Attention Fusion Module (AFM) concatenates channel attention and spatial attention, fusing the output features of the two branches. The essence of medical image segmentation tasks lies in locating the boundaries of objects in images. The Boundary Enhancement Module (BEM) is designed in the shallow layers of the network, focusing more on pixel-level edge details. Ou et al. [11] proposed a novel encoder-decoder visual Transformer architecture, Patcher, for medical image segmentation. Unlike standard visual Transformers, it utilizes patcher blocks to segment images into large blocks, each of which is further divided into small patches. Transformers are applied to the small patches within the large blocks, limiting the receptive field of each pixel. We intentionally let the large patches overlap to enhance communication within patches. The encoder adopts cascaded patch blocks with increasing receptive fields to extract features from local to global levels. This design enables Patcher to benefit from common coarse-to-fine feature extraction in CNNs and superior spatial relationship modeling in Transformers. We also propose a new decoder based on Mixture of Experts (MoE), which treats feature maps from the encoder as experts and selects an appropriate combination of expert features to predict labels for each pixel. However, these models, due to the introduction of Transformers, exhibit good

network performance, but the real-time performance of the entire network needs to be improved due to model complexity.

2.3 Lightweight Network with Medical Images

Subsequently, scholars have proposed methods to reduce the model size to ensure lightweight models meet the real-time requirements of clinical surgery [12]. AIADI proposes a single-layer unsupervised lightweight ear print recognition network that uses convolutional neural networks (CNN) and principal component analysis (PCA) for ear recognition [13]. The main novelty of MDFNet is its simple architecture, effectiveness, good trade-off between processing time and performance, and high robustness to occlusion. Chen et al. proposed a new lightweight network designed specifically for skin disease image segmentation [14], aiming to significantly reduce the number of parameters and floating-point operations while ensuring segmentation performance. Its designed ConvStem module features full-dimensional attention, learning complementary attention weights in all four dimensions of the convolution kernel, effectively enhancing the recognition of irregularly shaped lesion areas, reducing the number of model parameters and computational complexity, thereby promoting model lightweighting and performance improvement. The SCF Block reduces feature redundancy through spatial and channel feature fusion, significantly reducing the number of parameters while improving segmentation results. However, this method exhibits noticeable decreases in parameter and computational efficiency, and still lags behind EGEUnet in terms of parameter count. Additionally, the limited dataset and model generalization are areas for further research.

In summary, the technical gap for endoscopic images lies in finding an approach that can accurately and efficiently perform computational inference on medical images or videos with characteristics such as low texture, uneven illumination, and limited data volume, while also meeting the real-time requirements of clinical surgery and the constraints of computational resources on endoscopic devices. Therefore, we propose a lightweight feature extraction and matching network for endoscopic images based on adaptive attention, which efficiently extracts and matches features to meet the requirements of downstream tasks and clinical surgery.

3 Cross-Channel Multimodel Multistage Cascade Attention Network

Endoscopic images typically require real-time or fast feature extraction and matching to support surgical navigation and real-time decision-making. Therefore, the computational efficiency of algorithms must be prioritized. Inspired by the lightweight cascading group attention convolutional neural network EfficientViT, we designed a Cross-Channel Multimodal Multistage Cascade Attention Network (CMMCAN). CMMCAN is a lightweight network that can adaptively cascade attention across different modalities and stages under various constraints, enabling it to maintain high performance while accommodating fewer computational resources. The architecture overview is shown in Fig. 1.

The network architecture we have designed is based on an encoder-decoder framework at its core. During the encoding stage, the input undergoes preprocessing steps such as Downsampling, overlap embedding, and initialization, collectively referred to as the DOI module. Subsequently, we introduce our cross-modal adaptive attention module, ASFF, wherein twin network branches share weights to effectively capture both appearance and motion cues, enhancing scene comprehension. Following feature extraction at each layer, both branches feed into the SCIM module to extract multimodal complementary information and facilitate bidirectional transmission. Each branch adopts a convolutional neural network (CNN) architecture with distinct depths and structures to accommodate

diverse input modalities. Leveraging normalization and residual connections, coupled with self-adapting cross-modal attention via ASFF, accelerates model convergence. Moreover, the GCGFA module, positioned between encoders and decoders, orchestrates the aggregation of global context features through self-attention and cross-attention mechanisms, facilitating fusion and integration across different hierarchical levels and feature maps. Adaptive attention weights are then employed to weigh features from each location, ensuring robust performance in non-rigid environments, low-texture scenarios, and under uneven lighting conditions, typical of surgical navigation scenes often plagued by noise and blur.

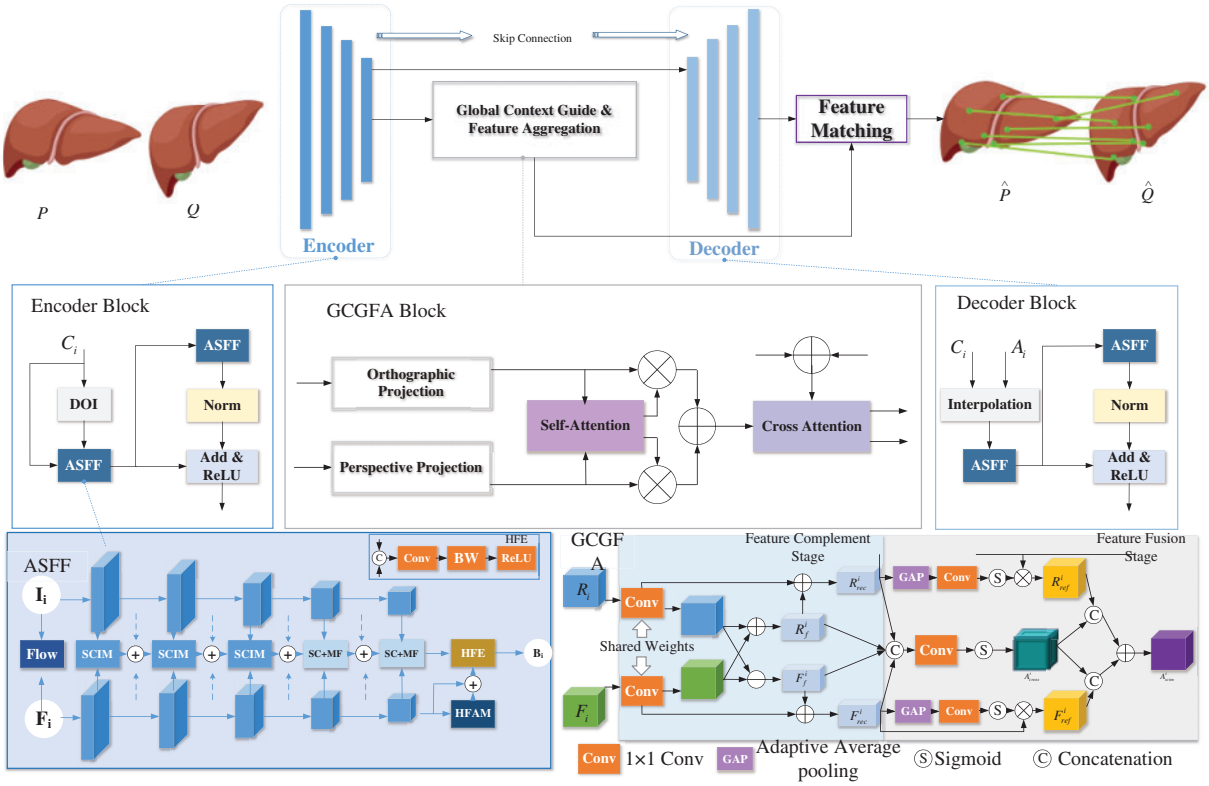


Figure 1: Network architecture with all pipeline

Our proposed lightweight network features adaptive attention, with the algorithm balancing global and local attention to enhance feature point extraction accuracy and density in scenarios with limited data quantity and quality. For details of the network framework design, refer to Section 3.1 Method Overview. Preliminary work and preprocessing details are provided in Section 3.2 Pre-Processing. The implementation strategy of the lightweight network with adaptive attention involves redesigning convolutional kernel sizes and quantities, adjusting convolutional network structures, designing lightweight modules, executing bitwise shifting and negation operations to replace multiplication in convolutions, reducing computational redundancy, and increasing computational speed to ensure real-time performance in clinical surgical scenarios. We propose cross-modal, cross-channel attention for global matching, introducing spatial attention for global matching of image features to address the limitations of quality prioritization and small data volume in monocular medical endoscopic images, enabling the network to achieve adaptive attention weight. Technical details of the encoder and decoder are provided in Section 3.3 Encoder-Decoder Architecture for Lightweight

Networks. Our proposed cross-modal adaptive attention module ASFF is detailed in Section 3.3.2 Multi-Modal Adaptive Self-Attention Feature Fusion Module–ASFF. The lightweight global context guidance and feature aggregation module are detailed in Section 3.3.3 SCIM, MFEM, HFAM, HFE Module. Additionally, the lightweight sparse optical flow module, overcoming inconsistencies in illumination and weak texture features, is detailed in Section 3.3.4 Location Search and Optical Flow Regression, facilitating keypoint detection and tracking to compensate for movement and deformation issues in limited data. For details on loss function design and other parameters and technical details, refer to Section 3.4 Loss Formulation.

3.1 Method Overview

The lightweight feature extraction and matching network for monocular endoscopic medical images based on adaptive attention, as proposed in this paper, is illustrated in Fig. 1. The network is built upon a classic encoder-decoder structure. We introduce an adaptive attention module based on cross-channel attention into a simple lightweight contrastive learning framework, along with a lightweight optical flow tracking module based on bottleneck structures to fuse multimodal image feature information. In the encoder, the input image's single channel is first downsampled, overlap-patch embedded, and initialized. It then enters our designed Multi-Modal Adaptive Self-Attention Feature Fusion Module (ASFF) to address inconsistencies in illumination and weak texture features for keypoint detection and tracking. This module leverages information about dynamic object motion from the optical flow map and appearance/color information from the RGB image. The twin branches within the ASFF module share weights, allowing them to simultaneously capture appearance and motion information for a more comprehensive scene understanding. Subsequently, in the decoder, interpolation is performed followed by decoding. In between, we design a global context feature guidance and aggregation module to compensate for the priority of medical endoscopic image quality and small data volume. Using orthogonal projection methods simplifies feature extraction calculations without considering factors like illumination, shadows, or distortions, while introducing perspective projection enhances spatial distribution feature calculations. By utilizing the uncorrelated relationship of features and introducing self-attention and cross-attention for weighting, both global and differential features can be effectively extracted in real-time and densified feature extraction and matching can be achieved.

Cross-modal, cross-channel attention is employed for global matching, while spatial attention is introduced for global matching of image features. Lightweight sparse optical flow features are contrastively learned with RGB features across modalities, utilizing different channels and optical flow signals for self-supervised learning of different views and features in the input image, extracting global and local features across channels and modalities, and simultaneously performing keypoint detection and tracking. Spatial attention is introduced to focus on specific regions of interest within the image. Cross-stage attention is introduced to facilitate the sharing of attention weights and feature information between different layers of the network.

3.2 Pre-Processing

Preprocessing: Endoscopic images are often affected by issues such as illumination, noise, and blur, thus requiring image quality assessment and preprocessing. Firstly, medical endoscopic image data undergo preprocessing, including image segmentation, scaling, normalization, and extraction of features from regions of interest (such as lesion areas). This includes denoising, contrast enhancement, and mitigating illumination changes.

Downsampling: To make the model more robust to input variations and increase the receptive field of each pixel, enabling the network to capture global information better, the input image sequences are uniformly downsampled to 512 × 512 size images. Experiments are also conducted with images at 256 × 256 and 128 × 128 sizes as control groups.

Overlap Embedded: By partitioning the input data into overlapping blocks, each block can capture local feature information, enhancing translational invariance and reducing the number of parameters. This approach also aids in capturing input feature information more comprehensively in subsequent self-attention mechanisms.

Initialization: Once all blocks are embedded into the feature vector space, initialization operations can be performed on these features. Initialization typically involves assigning initial weights or parameters to each dimension in the feature vector space, to be adjusted during subsequent training to perform specific tasks such as feature matching, object detection, etc.

Thus, block embedding operations are first performed to prepare the feature representation of the image, followed by initialization on these features. Subsequent training operations may then follow to fine-tune and learn network parameters to adapt to the specific task requirements.

3.3 Encoder-Decoder Architecture for Lightweight Networks

The implementation approach of the lightweight network with adaptive attention involves redesigning the size and number of convolution kernels, adjusting the convolutional network structure, designing lightweight modules, and performing bit shift and negation operations instead of multiplication in convolutions to reduce computational redundancy and improve processing speed, ensuring real-time performance in clinical surgical scenarios. We propose cross-modal, cross-channel attention for global matching, introducing spatial attention for global matching of image features, addressing the priority of quality in monocular endoscopic images and the limitation of small data volumes, enabling the network to adaptively adjust attention weights.

3.3.1 Siamese Network Framework

The lightweight Siamese network is a neural network architecture designed to measure the similarity between two input data pairs by comparing their feature representations. It is primarily used for metric learning and similarity comparison tasks. Based on the classical Siamese network structure, we designed a contrastive network structure based on RGB and optical flow (FLOW), which calculates the distance and correlation between RGB and FLOW feature vectors. This design is suitable for small datasets such as those found in medical scenarios and offers good interpretability. The output feature representations of these two subnetworks will be used to measure the similarity between input data pairs. Through maximizing the similarity of positive sample pairs and minimizing the similarity of negative sample pairs, the network learns in a self-supervised manner. The classical Siamese network structure is illustrated in Fig. 2a. Inputs x_1, x_2 are fed into two branches of the subnetwork, and the outputs E_w are obtained through the loss function. G_w represents an inference calculation comparing two branch networks in a network. When the loss function adopts the L_2 distance metric, the output of the network can be represented as:

$$E_w(X_1, X_2) = ||G_w(X_1) - G_w(X_2)|| \tag{1}$$

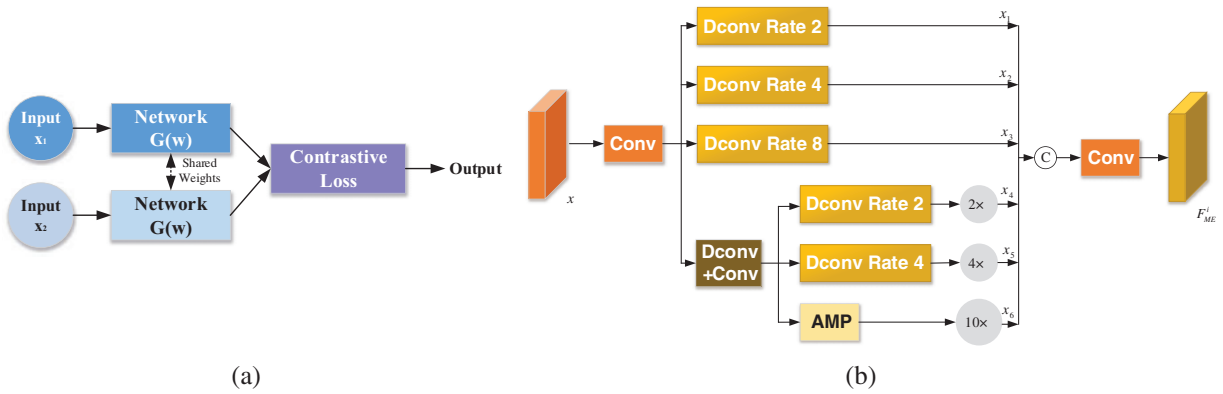


Figure 2: The Siamese network architecture and MFEM module. (a) Classic siamese network architecture, (b) MFEM module

The lightweight Siamese network framework we designed is an architecture composed of two identical five-layer subnetworks, as shown in Fig. 1. Each subnetwork receives an input data pair, and each layer at different levels has its own unique features to extract.

During the inference phase, the proposed framework is formulated as a local single-shot detection task. By precomputing the template branches of the Siamese subnetworks, the relevant layer representations are expressed as trivial convolution layers (Identity Convolution Layer) for online tracking, without introducing new feature transformations, but simply passing the input data directly to the output, serving to maintain the feature maps. This effectively increases the depth of the network without introducing additional non-linear transformations. The inputs to the two Siamese subnetworks are the single-channel information I_i processed through Downsampling, Overlap PatchEmbeded, and Initialization, and the inferred flow map F_i, which are also inputs to the ASFF module. The weights to be shared by the two subnetworks are first adaptively adjusted by the SCIM (Self-Adaptive Cross-Modal Interaction Module) module, and finally, the last two layers, in addition to passing through the SCIM module, also enter the MFEM (Multiscale Feature Enhancement Module) module [15], which encourages complementary information fusion between different modalities and different levels. The SCIM module and the MFEM module are illustrated in Figs. 1 and 2, respectively. We employ SCIM five times because different levels have their unique feature information to extract. Lower layers contain spatial local information, while higher layers contain semantic global information, both crucial for subsequent fusion stages. MFEM is used twice, enhancing the feature representation in medical endoscopy and surgical navigation scenarios by introducing multiscale feature enhancement and depth separable convolution. The joint use of these two modules can enhance the diversity of features fed into the attention heads, reducing redundancy in the attention map, ensuring feature extraction accuracy, and minimizing the number of parameters as much as possible.

3.3.2 Multi-Modal Adaptive Self-Attention Feature Fusion Module—ASFF

The flow map provides information about the dynamic motion of objects, while the RGB image provides appearance and color information. The twin branches in the ASFF module share weights, enabling them to simultaneously capture appearance and motion information, thereby comprehensively understanding the scene. After each branch undergoes feature extraction, they enter the SCIM module separately to extract multimodal complementary information. Each branch adopts

a convolutional neural network (CNN) architecture, and the features of the two branches have different depths and structures to adapt to different input modalities.

The ASFF module we designed is based on Siamese networks, incorporating five feature extraction layers for RGB and Flow. Depth separable convolution is used to reduce the network structure and capture different hierarchical features. An Adaptive Attention Module (SCIM) is added after each feature extraction layer for adaptive adjustment of window, stride, and neighborhood range. Here, features weighted by adaptive attention are combined with flow features. This allows the model to adjust the importance of features based on the similarity between positive and negative samples determined by attention-weighted features. Finally, the last two layers introduce MFEM to enhance the feature representation in medical endoscopy and surgical navigation scenarios by introducing multiscale feature enhancement and depth separable convolution, ensuring feature extraction accuracy while minimizing the number of parameters.

To restore spatial details and fully utilize global context information, an HFAM module is added after the five layers of flow feature extraction to recover some lost detailed information under the guidance of high-level semantic information. The features mapped and weighted by attention are multiplied after the fusion of RGB and flow through HFE, yielding attention-adjusted features for subsequent tasks. Matching and information fusion between RGB images and flow maps, input into the twin branches of the Siamese network, typically require the design of appropriate network architectures and training strategies. Table 1 summarizes the key components and specific operational steps of the Siamese network. For detailed calculations, please refer to Section 3.3.3.

Table 1: Key components and steps of the siamese network architecture

Step	Description
1	Input RGB image and optical flow
2	Extract features from RGB image and optical flow using feature extractors
3	Fuse the RGB and optical flow features
4	Process the fused features through the task-specific layer
5	Calculate the loss between the network output and the true labels
6	Perform backward propagation using the loss value and update the network parameters
7	Output the network predictions
8	Evaluate the network performance using a validation or test set

3.3.3 SCIM, MFEM, HFAM, HFE Module

The SCIM module can focus on information differences in each local area of the RGB and Flow graphs, filter out redundant information, and select more prominent feature representations. Taking the first layer as an example, the two Siamese network branch inputs are respectively I_i and F_i. First, the adjusted and supplemented RGB and Flow feature extraction are carried out I^i_{rec}, F^i_{rec}, respectively.

$$I^i_{rec} = (Conv_{1\times1}(I_i)) \oplus (Conv_{1\times1}(F_i) \otimes Conv_{1\times1}(I_i)) \tag{2}$$

$$F^i_{rec} = (Conv_{1\times1}(F_i)) \oplus (Conv_{1\times1}(F_i) - Conv_{1\times1}(I_i)) \tag{3}$$

where, $Conv_{1\times1}$ denotes 1×1 convolution layer. Then the fusion operation is carried out to further optimize the two parallel branches and supplement the detailed information. Specifically, this is achieved by performing global average pooling along the RGB and Flow channel directions, reducing the amount of computation by reducing the size of the feature map. The pooling follows the 1×1 convolution, Sigmoid activation function and is multiplied by itself. The refined sum is expressed as I^i_{ref} and F^i_{ref}:

$$I^i_{ref} = \sigma\left(Conv_{1\times1}\left(GAP\left(I^i_{rec}\right)\right)\right) \oplus I^i_{rec} \tag{4}$$

$$F^i_{ref} = \sigma\left(Conv_{1\times1}\left(GAP\left(F^i_{rec}\right)\right)\right) \oplus F^i_{rec} \tag{5}$$

GAP denotes an adaptive average pooling operation. The output of the SCIM module is expressed as:

$$A^i_{scim} = Cat\left[A^i_{cross}, I^i_{ref}\right] \oplus Cat\left[A^i_{cross}, F^i_{ref}\right] \tag{6}$$

where, $Cat\,[.]$ represents the concatenation operation, σ is the sigmoid activation function. The selected features and complementary features are generated to achieve cross-modal integration [16].

While adjusting adaptive attention weights across modes, lightweight networks are not as powerful as large-scale models in terms of feature extraction and information depth. In order to make more efficient use of Feature information of various scales, a Multiscale Feature Enhancement Module (MFEM) module was introduced into the last two layers of the ASFF module, as shown in Fig. 2. The MFEM module helps to encourage the fusion of complementary information between different modes and between different levels. The MFEM input first passes through the 1×1 convolutional layer, and then introduces the depth separable convolution of the common lightweight network design, reducing the number of model parameters and computational effort, while maintaining high feature extraction performance. This process can be explained by the following expression:

$$x_i = Conv^{up}_{1\times1}\left(Conv^{2^i}_{DW}(x)\right), (i = 1, 2, 3) \tag{7}$$

$$x_j = \times 2^{(j-3)}\left(Conv^{(j-3)}_{DW}\left(Conv^{dn}_{1\times1}\left(Conv_{DW}\left(Conv^{up}_{1\times1}(x)\right)\right)\right)\right), (j = 4, 5) \tag{8}$$

$$x_6 = \times 10\left(AMP\left(Conv^{dn}_{1\times1}\left(Conv_{DW}\left(Conv^{up}_{1\times1}(x)\right)\right)\right)\right) \tag{9}$$

$$M^i_{ME} = Conv^{dn}_{1\times1}\left(Cat\left[x_1, x_2, x_3, x_4, x_5, x_6\right]\right) \tag{10}$$

where, $Conv_{DW}$ represents discrete separable convolution, i represents dialation rate; $Conv^{up}_{1\times1}$ represents a 1×1 convolution for channel increment; $Conv^{dn}_{1\times1}$ denotes 1×1 convolution for channel decrement; and $\times2$, $\times4$, and $\times10$ denote the different upsampling multiples; $AMP\,(\cdot)$ denotes average maxpooling.

As we all know, the high-level features that guide the whole world have rich semantic information. As the width and depth of the network model increase, some detailed features will be lost, and the actual sensitivity field in convolution calculation is smaller than the theoretical sensitivity field. Therefore, in order to recover spatial details and make full use of global context information, this paper introduces HFAM [16] in the ASFF fusion feature layer. Specifically, the last two layers of high-level semantic features through SCIM are simply merged into global context guidance B_i. Since the low-level detail features vary according to scale, we first adaptively adjust the size to fit the low-level detail features, which helps to reduce the number of parameters. We first along the channel dimension connection high-level semantics and detail characteristics, and through the global average pooling

will change $R \times C \times H \times W$ into $R \times C \times 1 \times 1$. We then use the full connection layer to reduce the channel to C/r and then use the Re*lu* function. After repeating these two operations, the final weight is obtained, and the weight changes according to the characteristics of different details. HFAM is used to recover some lost details under the guidance of advanced semantic information. The specific calculation method is as follows:

$$H_i = \times 2\,(FC\&\mathrm{Re}lu)\,(GAP\,(Cat\,([AU\,(G)]\,, Conv_{1\times1}\,(Fi)))) \otimes AU\,(G) \oplus Conv_{1\times1}\,(Fi) \tag{11}$$

AU denotes adaptive upsampling through which the high-level semantic features are adjusted to the size of the detail features, $\times 2$ represents two repeated fully connected, Re*lu* represents activation function, *GAP* indicates adaptive average pooling.

3.3.4 Location Search and Optical Flow Regression

To address the issue of photometric changes, we introduce the Cross-Positioning Downsampling (CPD) and Optical Flow Regression (OFR) modules within the ASFF module. These modules facilitate the sharing of attention weights and feature information between different levels of the network's lateral connections. The Optical Flow Regression consists of simple convolutional layers, an optical flow estimation layer, and an optical flow regression layer. The optical flow estimation layer computes the optical flow between two consecutive frames, which is then regressed through the optical flow regression layer to predict the movement of key feature points between different frames. Dense optical flow estimation attempts to estimate the optical flow vector for every pixel in the image, whereas sparse optical flow estimation selects only a specific set of pixels for estimation.

Therefore, we introduce the lightweight model LiteFlowNet [17] to assist in optimizing RGB attention weights. LiteFlowNet aims to reduce the model's size by employing techniques such as depth separable convolution, making it suitable for resource-constrained environments, particularly for real-time optical flow estimation on mobile devices and embedded systems, such as mobile robots, autonomous vehicles, and drones.

3.4 Loss Formulation

Combining network architecture, task types and data characteristics, we design a loss function consisting of adaptive attention, location search and optical flow regression. Adaptive attention loss: The output of the adaptive attention module is the attention weight A, and the true label is A_{true}. We use cross entropy loss to measure the difference between them:

$$Loss_{attention} = CrossEntropyLoss\,(A, A_{true}) \tag{12}$$

Position search loss: Assume that the output of the position search module is the predicted position coordinates P, and the true position is P_{true}. We use the mean square error loss to measure the difference between them:

$$Loss_{loc} = \frac{1}{N}\sum_{i=1}^{N} \|P_i - P_{truei}\|_2^2 \tag{13}$$

Optical flow regression loss: Considering the possibility of deformation and noise in the endoscopic image, we use smoothness loss to estimate that the optical flow estimate is smooth between neighboring pixels. Huber Loss can reduce sensitivity to outliers and is suitable for noise or uncertainty that may be present in medical images.

$$Loss_{flow} = \begin{cases} 0.5\,(F - F_{true})^2, if\ |F - F_{true}| \le \delta \\ \delta\,|F - F_{true}| - 0.5\delta^2, if\ |F - F_{true}| > \delta \end{cases} \tag{14}$$

where, F is the predicted optical flow in the output of the optical flow regression task F_{true} is the real optical flow. In summary, the synthetic loss function can be the weighted sum of these sub-loss functions:

$$TotalLoss = \omega_{attention} Loss_{attention} + \omega_{loc} Loss_{loc} + \omega_{flow} Loss_{flow} \tag{15}$$

$\omega_{attention}$, ω_{loc} and ω_{flow} are the weights for each subtask to balance the importance of different tasks, adjusted along with the network training process to balance the importance of different tasks in model reasoning. In the overall network, this comprehensive loss function is used to guide the network through training to determine the best parameter and weight configuration so that the network can achieve the goal of feature extraction and matching.

4 Experiments

4.1 Dataset and Implementation Details

4.1.1 Experimental Environment and Datasets

The feature extraction and matching method of monocular medical endoscope image based on adaptive attention mechanism is proposed in this paper. With the help of deep learning framework PyTorch, the corresponding calculation and training process are realized by using the function of appearance calculation and automatic differentiation. We employed the PyTorch framework to construct the network model, which was accelerated on a single NVIDIA RTX 4080 GPU with 12 GB of video memory, and equipped with a Gen Inter Core i9-13900H CPU, ensuring sufficient processing power for the learning algorithms. This comprehensive experimental setup ensured a robust and reliable environment for conducting the experiments and obtaining accurate results.

Furthermore, in our medical applications, we adhere strictly to medical standards and privacy regulations to ensure that the processing and storage of image data comply with all legal requirements. The datasets utilized in this study are outlined in Table 2. For model training and cross-validation, we employed medical datasets containing endoscopic images, including HyperKvasir, EAD, M2caiseg, CVC-ClinicDB, Kvasir-SEG, and UCL synthetic datasets. Additionally, clinical endoscopic videos and CT data from 83 anonymous patients from the Tianjin Academy of Traditional Chinese Medical Affiliated Hospital (all data approved for use and all anonymized) were also included, collectively comprising the experimental dataset used for model training and evaluation. The first line of Fig. 5 shows a representative sample of the dataset. During the training phase, we utilized adjacent frames from each video with varying intervals. To strike a balance between overlap and spacing between frame pairs, three intervals were selected for experimentation: 1, 4, and 16. Consequently, the final dataset consisted of 8352 training pairs and 2217 testing pairs. Both testing and training input frames were resized to 256 × 256 pixels to ensure consistency and compatibility with the model architecture.

4.1.2 Model Training and Hyperparameters

In this paper, the training process utilized a batch size of 16 and executed 300,000 iterations. An Adam optimizer with a learning rate of 0.00002 was employed for optimization. During the warm-up phase of network training, the learning rate gradually increased from 6% to 100% of its standard value over 8000 training epochs. Subsequently, the learning rate was maintained at 0.0008 for 180,000 epochs before gradually decreasing to 5% of its standard value. The model input size was 256 × 256, and data

augmentation techniques were applied to enhance the robustness and generalization capabilities of the model. Given the utilization of a novel siamese network model in lieu of the mainstream cascade attention-based EfficientViT lightweight model, we evaluated the model comprehensively in terms of model parameters, memory footprint per execution, floating-point operations (FLOPs), model size, and runtime. Furthermore, we conducted a comparative analysis with lightweight models released by MIT and the Chinese University of Hong Kong, namely EfficientViT, ShuffleNet, GCGLNet, FMMI, and DenseDescriptor. The training results of different methods on the six listed datasets are summarized in Table 5. According to the model training process, the optimal hyperparameter configuration during our training is shown in Table 3.

Table 2: Datasets used in the study

Dataset name	Data type	Dataset size	Train size	Test size
Kvasir-SEG [18]	Colonoscopy images	Approx. 1,1 million images and video frames	100,000	20,000
EAD Datasets [19]	Endoscope surgery images	11552 Images (2002–2021)	8710	1742
M2caiSeg [20]	Endoscope images, CT, MRI	Approx. 307 images	260	47
CVC-ClinicDB [21]	Colonoscopy images	Approx. 612 images	500	112
UCL Synthetic Dataset [22]	Synthetic image data	Varies by task	260	47
Clinical endoscopic video and CT data from 83 anonymous patients	Endoscope surgery Images & CT	A large and diverse sample of images	70	13

Table 3: Initially hyperparameters setup

Batch size	β_1	β_2	Learning rate	Epoch-1	Epoch-2	α	γ	Linear increase
4/8	0.9	0.999	2e-5	70	10	0.0001	0.9	0.005

4.2 Homography Estimation and Loss/Accuracy Curve

For each matching pair of points (x, x'), where x denotes the coordinates of a point in one image and x' represents the corresponding coordinates of that point in the other image, the homography matrix H can be solved using the following linear system: $x' = Hx$. Following the EfficientViT model, we report the precision and recall rates for matching encounters with GT, with a reprojection error threshold of 3 pixels. Additionally, we evaluate the accuracy of estimated homographies from the correspondences using both robust and non-robust solvers: Random Sample Consensus (RANSAC) and weighted Direct Linear Transformation (DLT) [23]. For each image pair, we compute the average reprojection error at the four image corners and report the area under the cumulative error curve (AUC) for thresholds up to 1 px and 5 px. In alignment with best practices for benchmarking, unlike past works [24,25], we employ state-of-the-art robust estimators and conduct extensive threshold tuning for each method individually. When compared to classical sparse feature extraction and matching methods [26,27], we report the highest-scoring results, with quantitative computations summarized in Table 4. The findings indicate that CMMCAN can generate superior correspondences,

exhibiting the highest precision (P) and recall (R). Consequently, this leads to more accurate results when estimating homographies using both RANSAC and even the faster least-squares solver, DLT.

Table 4: CMMCAN homography estimation. R-high recall, P-highest precision

Features + Matcher		R↑					P↑					AUC-RANSAC		AUC-DLT↑	
		Orin	SD20	SD45	$K3^2$	$K5^2$	Orin	SD20	SD45	$K3^2$	$K5^2$	@1 px	@5 px	@1 px	@5 px
Dense	LoFTR	–	–	–	–	–	92.7	92.5	92.4	92.5	**92.0**	41	78	38	70
SuperPoint	Sparse 1	72.1	72.0	71.7	71.9	71.7	68.8	68.7	68.6	68.4	67.9	35.1	75.7	32.4	20.8
	Sparse 2	98.2	98.0	98.3	98.0	98.4	81.4	81.2	81.0	81.3	81.2	37.4	76.4	33.8	76.9
	Sparse 3	**99.1**	99.1	**95.0**	**94.9**	**94.7**	83.6	83.4	83.3	83.5	83.2	**38.6**	79.0	34.6	77.9
	CMMCAN	99.8	**99.7**	94.6	94.2	93.6	**89.2**	**89.1**	**89.1**	**89.0**	**88.8**	38.3	**79.6**	**35.1**	**78.5**

In addition, in order to verify the robustness of the model to noise and fuzzy, based on the test data set, random sampling divided the test set into two parts, and added Gaussian noise with standard deviation of 20 and 45 respectively to generate a new disturbed data set with noise, as shown in Fig. 3. The perturbation data set is used to test the CMMCAN model and calculate the performance index under noise interference. It is listed in columns SD20 and SD45 of the Fig. 3.

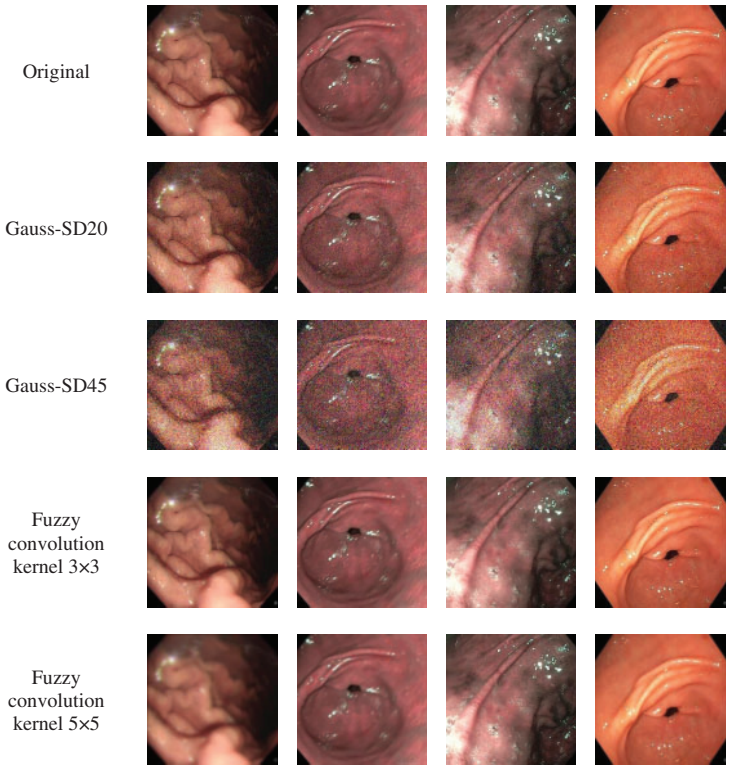
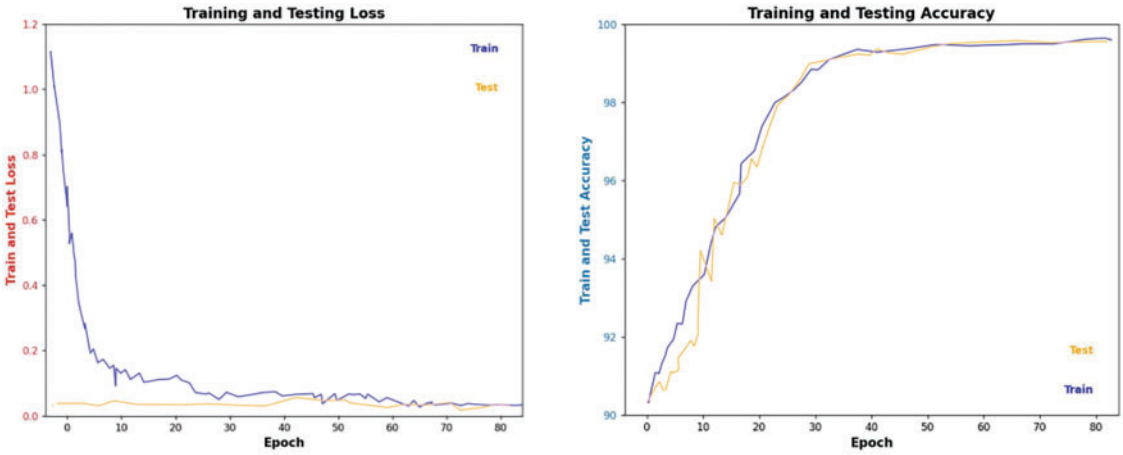


Figure 3: Examples of varying levels of noise and blur to validate the robustness of CMMCAN, (from the EAD dataset). The first row showcases original image examples, while the second row depicts images with added Gaussian noise with a standard deviation of 20. The third row displays images with added Gaussian noise with a standard deviation of 45. The fourth row illustrates examples of images blurred using a 3×3 convolutional kernel, while the fifth row shows examples of images blurred using a 5×5 convolutional kernel

Furthermore, to evaluate the model's robustness to noise and blur, we partitioned the test dataset into two subsets by randomly sampling, each augmented with Gaussian noise of standard deviation 20 and 45, respectively, generating new noisy perturbed datasets as illustrated in Fig. 3. We tested the CMMCAN model using these perturbed datasets and computed performance metrics under noise interference, listed in the columns SD20 and SD45 of Table 4.

To assess the model's resistance to disturbances, we introduced motion blur into the test dataset using convolution operations typically associated with moving objects in a consistent direction. The blur kernel angle was set randomly, simulating various degrees of blur by adjusting the level of rotation matrix. We applied 3×3 and 5×5 convolution kernels for different levels of blur, where larger kernels corresponded to higher levels of blur and smaller kernels led to more localized blur. Convolving the original clear images with this series of blur kernels, we generated corresponding blurry endoscopic image datasets as depicted in the fourth and fifth lines of Fig. 4. The CMMCAN model was evaluated using these blurred datasets, and performance metrics under random degree of blur were listed in columns K32 and K52 of Table 4, respectively.



(a) (b)

Figure 4: Loss (a) and accuracy (b) curve of the model during the model inference process with Epoch 80 on the EAD original map

Experimental results demonstrate that CMMCAN can establish better correspondences, exhibiting the highest precision (P) and recall (R). Therefore, this leads to more accurate results when using RANSAC or even faster least squares solver DLT to estimate homography. The introduction of different levels of noise had negligible impact on the performance of CMMCAN feature extraction and matching, with the recall rates decreasing by only 0.11% and 0.22% when introducing Gaussian noise with standard deviations of 20 and 45, respectively, and the highest precision decreasing by 0.63% and 0.22%. Introducing varying degrees of blur resulted in a decrease in recall rate by 0.11% and the highest precision by 0.34%. This indicates that different levels of noise and blur have minimal impact on the performance, and the accuracy is hardly affected. Thus, validating the proposed model's robustness to noise and blur.

4.3 Quantitative and Qualitative Comparison with Prior Work

4.3.1 Quantitative Comparison with Prior Work

We conducted a comprehensive comparison of GCGLNet against 6 state-of-the-art (SOTA) methods, including EfficientViT-MIT [28], EfficientViT-CUHK [29], ShuffleNet [30], GCGLNet [16], FMMI [31], and DenseDescriptor [32]. The first three methods were originally designed for image classification in constrained scenarios, emphasizing lightweight network models. GCGLNet and FMMI, on the other hand, focus on scene understanding based on multimodal information fusion. DenseDescriptor is capable of extracting dense feature points from endoscopic medical images. To ensure a fair comparison, we modified the input image sizes and conducted all computations on our platform, leveraging the reported descriptions and released codebases. To assess the quality of the extracted features, we use a simple random forest to classify organs of concern within the visual threshold. The quantitative results of different method models' inferences on the various datasets are summarized in Table 5.

Table 5: CMMCAN performance results on transfer learning datasets. Our model achieves new state-of-the-art accuracy for 6 out of 6 datasets, with 9.6× fewer parameters on average. Bold indicates the best results calculated with the same data set. Acc is the Top5 Set the Epoch to 60

	Methods	Acc↑	IoU↑	Precision/%↑	F1↑	Params/M↓	FLOPs/M↓
HyperKvasir	EfficientViT-MIT-B3	53.1	30.6	68.3	0.5563	49	124
	EfficientViT-CUHK-M4	48.8	34.9	62.1	0.5546	8.8	299
	ShuffleNet-V2	29.8	16.9	50.7	0.4768	**6.9**	427
	GCGLNet	60.8	55.6	78.5	0.6163	7.87	7209
	FMMI	64.2	59.0	86.2	0.6629	593	6264
	DenseDescriptor	**89.2**	**72.8**	**89.0**	**0.7463**	1365	12149
	Ours	85.6	68.5	86.1	0.7390	11.7	**89**
EAD	EfficientViT-MIT-B3	50.0	22.5	56.8	0.5454	49	135
	EfficientViT-CUHK-M4	51.7	36.2	61.4	0.4733	8.8	324
	ShuffleNet-V2	20.0	15.1	44.1	0.4587	**6.9**	398
	GCGLNet	61.2	43.6	79.8	0.6223	7.87	7382
	FMMI	70.8	58,1	78.3	0.6505	593	6478
	DenseDescriptor	**90.6**	73.8	**86.5**	**0.7889**	1365	13306
	Ours	88.4	**74.1**	82.7	0.7852	9.6	**92**
M2caiseg+ CVC-ClinicDB+Kvasir-SEG	EfficientViT-MIT-B3	46.7	10.9	43.5	0.4679	49	523
	EfficientViT-CUHK-M4	35.5	11.9	57.3	0.3567	8.8	517
	ShuffleNet-V2	12.7	0.0	34.4	0.3343	**6.9**	461
	GCGLNet	22.4	5.9	68.2	0.6346	7.87	8962
	FMMI	35.6	6.2	76.8	0.5166	593	8325
	DenseDescriptor	**65.4**	9.6	**82.1**	0.7421	1365	26703
	Ours	55.2	**12.7**	76.9	**0.7939**	11.7	**217**
UCLSynthetic Dataset	EfficientViT-MIT-B3	53.2	36.2	72.8	0.4465	49	132
	EfficientViT-CUHK-M4	57.7	45.5	74.3	0.5768	8.8	119
	ShuffleNet-V2	31.4	15.9	65.2	0.5697	**6.9**	310
	GCGLNet	67.2	51.2	78.9	0.7456	7.87	6450
	FMMI	73.0	62.8	95.5	0.7637	593	4621
	DenseDescriptor	**95.9**	**78.9**	**96.7**	**0.8058**	1365	102972
	Ours	89.2	73.2	93.9	0.7923	11.7	**58**

Our evaluation criteria encompassed various metrics that reflect the performance of each method in terms of accuracy, efficiency, and robustness. By comparing GCGLNet against these diverse SOTA approaches, we aimed to demonstrate its superiority in handling endoscopic medical image analysis tasks, particularly in scenarios that require multimodal information fusion and dense feature extraction.

From the perspective of feature extraction accuracy, the DenseDescriptor method demonstrated the highest accuracy in feature extraction. However, due to its larger model size, it requires more computational resources for dense feature extraction, resulting in longer processing time and thus limited efficiency. In contrast, our proposed CMMCAN method eliminates redundant computations and modules, enabling it to better utilize complementary information from multimodal descriptions without increasing computational overhead. This allows for more accurate and faster feature extraction and matching. Although CMMCAN may slightly lag behind DenseDescriptor in terms of absolute accuracy, it exhibits a significant advantage in feature extraction efficiency. Moreover, results demonstrate that compared to EfficientViT-B3 with cascading attention, the proposed algorithm achieves an 75.4% improvement in Acc, along with enhanced runtime performance and storage efficiency. Furthermore, the proposed approach improvement F1 score by 33.2% accelerates runtime by 70.2%, and achieves a better balance between model training cost, real-time performance, and precision.

4.3.2 Qualitative Comparison with Prior Work

The qualitative comparison results of different methods on the corresponding six datasets are presented in Fig. 5. These qualitative findings are generally consistent with the quantitative results shown in Table 5. The results indicate that CMMCAN outperforms the comparison methods in terms of efficiently extracting feature points. Notably, DenseDescriptor achieves the best performance in feature extraction, but due to its larger model size, it requires more time and computational resources, making it suitable for extracting dense feature points. Conversely, ShuffleNet-V2 extracts the fewest feature points. While this model has a smaller size, its cost-effectiveness for feature extraction in monocular endoscopy scenarios is relatively low.

In summary, CMMCAN offers the highest cost-effectiveness in feature extraction compared to the comparison methods. Specifically, it utilizes the most efficient model to extract the maximum number of accurate feature points. This advantage of CMMCAN makes it a promising approach for endoscopic image analysis tasks that require accurate and efficient feature extraction.

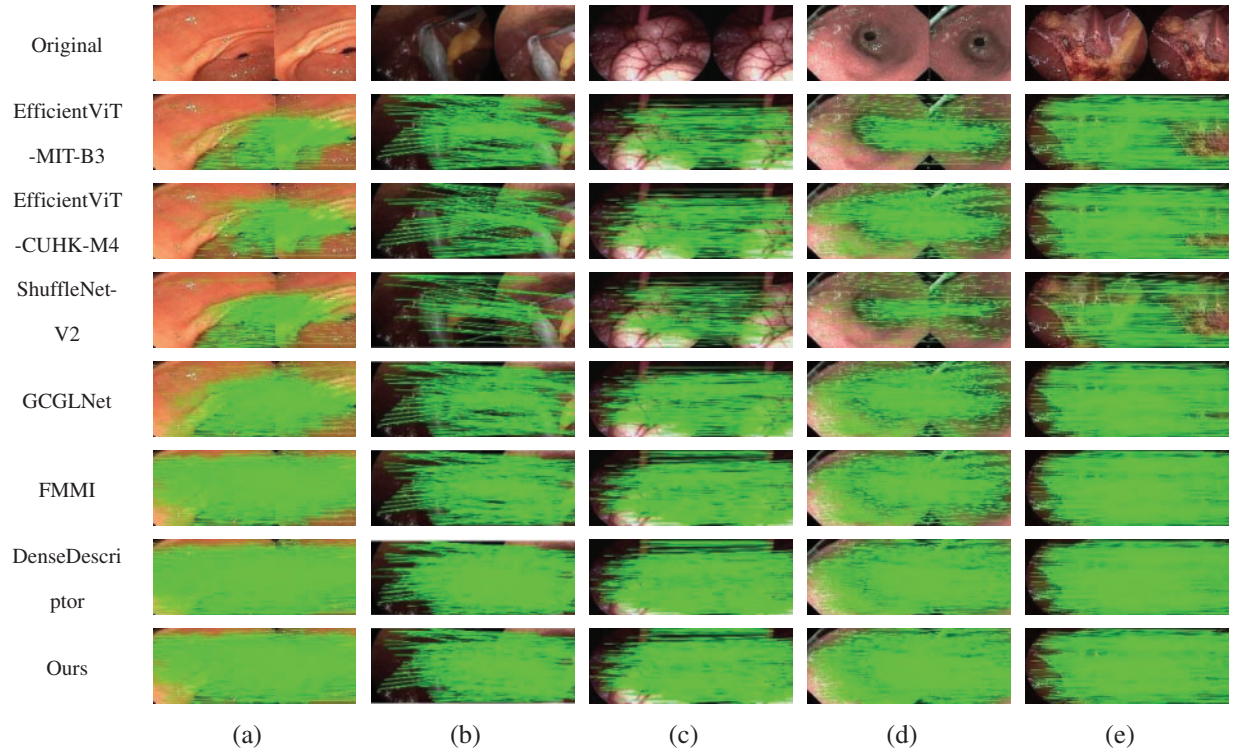


Figure 5: Comparison of qualitative results. The four columns on the left (a)–(d) are the four datasets corresponding to Table 5, and the last column (e) is the patient cases obtained from the hospital

4.4 Ablation Studies

In this subsection, we conducted ablation experiments to assess the performance of our algorithm under different settings of model modules and variables. Specifically, we conducted ablation studies on the cross-channel cascade attention module (CrossCAM), adaptive attention module (CAM), cross-modal cascade adaptive attention module (ASFF), and the number of cascade layers (N). Quantitative evaluation results were provided for each experiment. As primary evaluation metrics, we employed accuracy, precision, model parameter count, mean accuracy (mACC), mean intersection over union (mIoU), average runtime, and endpoint error (EPE) for optical flow analysis. Additionally, recall and precision were also considered. To present the performance more clearly, we visualized the results using confusion matrices and ROC curves. Through these ablation experiments, we aimed to gain insights into the individual contributions of each module and variable to the overall performance of our algorithm. This analysis allowed us to identify the most effective configurations and optimize the model for improved accuracy, efficiency, and generalizability. The ablation experimental results of the model on the HyperKvasir and EAD datasets are shown in Table 6.

Table 6: The ablation experimental results of the model on the HyperKvasir and EAD datasets. The symbol "✓" indicates the utilization of the module proposed in this study. "ALL" signifies the evaluation of all pixels, whereas "NOC" specifically refers to the testing conducted solely on pixels in non-occluded regions. The average results of the ablation experiments, including param (model parameters), precision, recall, and times (average runtime), are provided separately for the HyperKvasir and EAD datasets

Network setting				HyperKvasir		EAD		Param/M	Precision	Recall	Time (ms)
				D1-all (%)		3 px (%)					
CrossC	CAM	ASFF	N	NOC	ALL	NOC	ALL				
–	–	–	2	56.5	34.3	43.9	42.1	6.02	60.1	90.5	11.3
✓	–	–	2	52.1	58.2	56.2	47.8	7.33	62.6	92.1	42.5
✓	✓	–	2	68.2	64.5	69.7	63.7	9.77	68.9	92.6	46.3
✓	✓	✓	2	74.4	72.4	70.5	69.3	9.86	76.4	94.7	52.1
✓	✓	✓	3	84.1	82.9	81.0	80.2	10.05	80.5	96.2	68.7
✓	✓	✓	4	**89.6**	**87.0**	**88.2**	**85.3**	10.65	**84.4**	**97.0**	**89.5**

The ablation experiments tested the key modules of the model: cross attention, channel attention, ASFF, and the number of convolutional layers. The cross-attention module improved the model by 28%, CAM improved the model by 20.5% on average in the two data sets, and ASFF improved the model performance by 5.43%. The more convolutional layers, the more accurate the calculation, but the correspondingly longer the time. Therefore, the setting of the model and the number of layers should balance local computing resources and timeliness requirements to make appropriate cuts.

5 Conclusion and Discussion

To address challenges such as low texture, uneven illumination, and non-rigid structures in endoscopic surgical scenes, this study introduces a novel Cross-Channel Multi-Modal Adaptive Spatial Feature Fusion (ASFF) module based on the lightweight architecture of EfficientViT. Additionally, we propose a novel lightweight feature extraction and matching network, named Cross-Modal Multi-Channel Attention Network (CMMCAN). This network dynamically adjusts the attention weights for cross-modal features of different channels and optical flows within the contrastive learning framework of a dual-branch Siamese network. It integrates both static and dynamic information of endoscopic image sets to counteract common issues like texture degradation, lighting variations, and non-rigid structures. This ensures the robustness of feature extraction across various widths, noise levels, and blur scenarios.

While the proposed method achieves superior quantitative and qualitative results in comparative experiments, there are still areas for improvement. Firstly, the proposed model assumes that the observed data has low texture, uneven illumination, and non-rigid structures. Performance degradation may occur under non-assumed conditions, indicating the need for further research on model generalization. Additionally, medical endoscopic datasets are inherently limited in quantity and quality. Although we employ enhancement methods such as cropping, scaling, and normalization in preprocessing, richer datasets for model training may lead to better results. Finally, in clinical surgical scenarios, there is a particular emphasis on real-time and accurate AI tools. Therefore, optimizing model parameters and extracting denser features more efficiently to enhance downstream task performance and quality remains a focus of our ongoing efforts.

Acknowledgement: The authors would like to thank those who contributed to the article and who support them from Hebei University of Technology, Tianjin Renai College and Tianjin Academy of Traditional Chinese Medical Affiliated Hospital for the experience and technical support. We specifically consulted Kewei Wei from the Department of Orthopaedics, Tianjin Fifth Central Hospital expertise to validate our robustness experiments and ensure that our proposed method meets clinical practical requirements.

Funding Statement: This work was supported by Science and Technology Cooperation Special Project of Shijiazhuang (SJZZXA23005).

Author Contributions: Study conception and design: Nannan Chong, Fan Yang; data collection: Nanan Chong; analysis and interpretation of results: Nannan Chong; draft manuscript preparation: Nannan Chong, Fan Yang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Some data are available on request from the authors, which support the findings of this study are available from the corresponding author, Nannan Chong, upon reasonable request. Other data are not available due to patient privacy restrictions. All the data involved in the experiment came from publicly available datasets or were anonymized by hospitals and allowed to be used.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] K. Kaplan, Y. Kaya, M. Kuncan, and H. M. Ertunç, "Brain tumor classification using modified local binary patterns (LBP) feature extraction methods," *Med. Hypotheses*, vol. 139, no. 10, pp. 109696, 2020. doi: 10.1016/j.mehy.2020.109696.

[2] S. B. G. T. Babu and C. S. Rao, "Efficient detection of copy-move forgery using polar complex exponential transform and gradient direction pattern," *Multimed. Tools Appl.*, vol. 82, no. 7, pp. 10061–10075, 2023. doi: 10.1007/s11042-022-12311-6.

[3] Priyanka and D. Kumar, "Feature extraction and selection of kidney ultrasound images using GLCM and PCA," *Procedia Comput. Sci.*, vol. 167, no. 10, pp. 1722–1731, 2020. doi: 10.1016/j.procs.2020.03.382.

[4] S. Dua, J. Singh, and H. Parthasarathy, "Image forgery detection based on statistical features of block DCT coefficients," *Procedia Comput. Sci.*, vol. 171, no. 1, pp. 369–378, 2020. doi: 10.1016/j.procs.2020.04.038.

[5] B. Monika, K. Munish, and K. Manish, "2D object recognition: A comparative analysis of SIFT, SURF and ORB feature descriptors," *Multimed. Tools Appl.*, vol. 80, no. 12, pp. 18839–18857, 2020. doi: 10.1007/s11042-021-10646-0.

[6] T. K. Araghi and A. A. Manaf, "An enhanced hybrid image watermarking scheme for security of medical and non-medical images based on DWT and 2-D SVD," *Future Gener. Comput. Syst.*, vol. 101, no. 2, pp. 1223–1246, 2019. doi: 10.1016/j.future.2019.07.064.

[7] T. Shi, N. Boutry, Y. Xu, and T. Geraud, "Local intensity order transformation for robust curvilinear object segmentation," *IEEE Trans. Image Process.*, vol. 31, pp. 2557–2569, 2022. doi: 10.1109/TIP.2022.3155954.

[8] L. Wei *et al.*, "Quantum machine learning in medical image analysis: A survey," *Neurocomputing*, vol. 525, no. 5786, pp. 42–53, 2024. doi: 10.1016/j.neucom.2023.01.049.

[9] M. Liao, H. Tang, X. Li, P. Vijayakumar, V. Arya and B. B. Gupta, "A lightweight network for abdominal multi-organ segmentation based on multi-scale context fusion and dual self-attention," *Inf. Fusion*, vol. 108, no. 4, pp. 102401, 2024. doi: 10.1016/j.inffus.2024.102401.

[10] X. Fan, J. Zhou, X. Jiang, M. Xin, and L. Hou, "CSAP-UNet: Convolution and self-attention paralleling network for medical image segmentation with edge enhancement," *Comput. Biol. Med.*, vol. 172, no. 11, pp. 108265, 2024. doi: 10.1016/j.compbiomed.2024.108265.
[11] Y. Ou, J. Pan, J. Dai, Z. Sun, and Y. Xiao, "Patcher: Patch transformers with mixture of experts for precise medical image segmentation," presented at *MICCAI 2022*, Singapore, 2022, pp. 475–484.
[12] A. Oussama, B. Khaldi, and M. L. Kherfi, "A fast weighted multi-view Bayesian learning scheme with deep learning for text-based image retrieval from unlabeled galleries," *Multimed. Tools Appl.*, vol. 82, no. 7, pp. 10795–10812, 2023. doi: 10.1007/s11042-022-13788-x.
[13] O. Aiadi, B. Khaldi, and M. L. Kherfi, "MDFNet: An unsupervised lightweight network for ear print recognition," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 10, pp. 13773–13786, 2023. doi: 10.1007/s12652-022-04028-z.
[14] H. Chen, B. Khaldi, and C. Saadeddine, "SCSONet: Spatial-channel synergistic optimization net for skin lesion segmentation," *Front. Phys.*, vol. 12, pp. 1388364, 2024. doi: 10.3389/fphy.2024.1388364.
[15] T. Gong, W. Zhou, X. Qian, J. Lei, and L. Yu, "Global contextually guided lightweight network for RGB-thermal urban scene understanding," *Eng. Appl. Artif. Intell.*, vol. 117, no. 12, pp. 105510, 2023. doi: 10.1016/j.engappai.2022.105510.
[16] N. Huang, W. Zhou, X. Qian, J. Lei, and L. Yu, "Middle-level feature fusion for lightweight RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 31, pp. 6621–6634, 2022. doi: 10.1109/TIP.2022.3214092.
[17] T. W. Hui, Q. Jiao, Q. Zhang, and J. Han, "LiteFlowNet: A lightweight convolutional neural network for optical flow estimation," in *Proc. CVPR 2018*, Salt Lake City, UT, USA, 2018, pp. 8981–8989.
[18] A. R. Røe, M. Riegler, and P. Halvorsen, "KVASIR: A multi-class image dataset for computer aided gastrointestinal disease detection," in *Proc. 8th ACM Multimed. Syst. Conf.*, Taipei, Taiwan, 2013, pp. 164–169.
[19] S. Ali *et al.*, "Endoscopy artifact detection (EAD 2019) challenge dataset," arXiv preprint arXiv:1905.03209, 2019.
[20] S. Maqbool, M. Riegler, and P. Halvorsen, "m2caiSeg: Semantic segmentation of laparoscopic images using convolutional neural networks," arXiv preprint arXiv:2008.10134, 2020.
[21] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "DoubleU-Net: A deep convolutional neural network for medical image segmentation," in *Proc. CBMS*, Rochester, MN, USA, Jul. 2020, pp. 28–30.
[22] Y. Wang *et al.*, "UCL-Dehaze: Towards real-world image dehazing via unsupervised contrastive learning," *IEEE Trans. Image Process.*, vol. 33, pp. 1361–1374, 2024. doi: 10.1109/TIP.2024.3362153.
[23] F. Barone, M. Marrazzo, and C. Oton, "Camera calibration with weighted direct linear transformation and anisotropic uncertainties of image control points," *Sensors*, vol. 20, no. 4, pp. 1175, 2020. doi: 10.3390/s20041175.
[24] P. Gotardo, O. Bellon, and L. Silva, "Range image segmentation by surface extraction using an improved robust estimator," in *Proc. CVPR*, Madison, WI, USA, Jun. 18–20, 2003, vol. 2, pp. II–33.
[25] S. Ourselin, A. Roche, S. Prima, and N. Ayache, "Block matching: A general framework to improve robustness of rigid registration of medical images," in *Proc. MICCAI*, PA, USA, Oct. 11–14, 2000, pp. 557–566.
[26] K. Luo, T. Guan, L. Ju, H. Huang, and Y. Luo, "P-MVsNet: Learning patch-wise matching confidence aggregation for multi-view stereo," in *Proc. ICCV*, Seoul, Republic of Korea, Oct. 27–Nov. 2, 2019, pp. 10452–10461. doi: 10.1109/ICCV43118.2019.
[27] Y. Shi, T. Guan, L. Ju, H. Huang, and Y. Luo, "ClusterGNN: Cluster-based coarse-to-fine graph neural network for efficient feature matching," in *Proc. CVPR*, New Orleans, LA, USA, Jun. 19–24, 2022, pp. 12517–12526.
[28] H. Cai, J. Li, M. Hu, C. Gan, and S. Han, "EfficientViT: Lightweight multi-scale attention for on-device semantic segmentation," arXiv preprint arXiv:2205.14756, 2022.

[29] X. Liu, H. Peng, N. Zheng, and Y. Yang, "EfficientViT: Memory efficient vision transformer with cascaded group attention," in *Proc. CVPR*, Vancouver, BC, Canada, Jun. 18–22, 2023, pp. 14420–14430.
[30] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. CVPR*, Salt Lake City, UT, USA, Jun. 18–22, 2018, pp. 6848–6856.
[31] D. S. Khafaga *et al.*, "Hybrid dipper throated and grey wolf optimization for feature selection applied to life benchmark datasets," *Comput. Mater. Contin.*, vol. 74, no. 2, pp. 4531–4545, 2023. doi: 10.32604/cmc.2023.033042.
[32] X. Liu *et al.*, "Extremely dense point correspondences using a learned feature descriptor," in *Proc. CVPR*, Seattle, WA, USA, Jun. 2020, pp. 14–19.