Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25–29 May 2026, Vancouver, BC

Explainable Medical Image Segmentation via Attention-Gated Fusion of ViT and U-Nets

Mahmoud Khalaf[†,*], Robin Cohen[†], Saidul Islam [‡], Jamal Bentahar[‡]
[†] University of Waterloo
[‡] Concordia University

Abstract

Medical image segmentation is essential for assisting medical professionals in locating anomalies in images. The lack of explainability in current medical image segmentation frameworks demonstrates a gap in assisting clinicians in understanding how segmentation decisions are made, towards identifying the segmentation target. In this paper, we present a framework that offers an improved approach for assisting medical professionals in locating anomalies while providing visual explanations in the form of heatmaps of the target. We propose a dual encoder architecture using a U-Net encoder and ViT to perform accurate segmentation. We employ an attention fusion mechanism to fuse both encoder embeddings and generate an explainability heatmap that offers improved results for highlighting important features. We include discussion that reflects on the ways in which our approach advances the state of the art for medical decision making, in comparison with other current research, elaborating as well as on how the approach can be of value for distinct healthcare concerns. While our current results focus on how our dual encoder approach yields significant benefit, we also briefly discuss how to integrate textual explanations alongside, as a valued step forward for future work.

Keywords: Explainable AI, Medical Applications of AI, Computer Vision Segmentation, AI for Social Good, Transformers, Attention.

1. Introduction

Deep learning has significantly broadened the horizons for computer vision applications across various domains, with medical image analysis as one of the high-impact areas. In particular, medical image segmentation plays an important role in detecting anomalies and can save the lives of many patients. Deep learning models have achieved state of the art accuracy in segmentation tasks [1]; however, their black-box nature poses a significant challenge in clinical settings where interpretability and trust are imperative. Medical professionals require not only accurate segmentations, but also clear explanations of why target regions are delineated. Existing medical image segmentation approaches predominantly focus on maximizing segmentation accuracy metrics through innovations. However, they often neglect the aspect of explainability.

Among Convolutional Neural Networks (CNNs) architectures, U-Net [1] and its variants have demonstrated strong performance in medical image segmentation tasks, becoming the standard architecture due to their encoder-decoder structure and ability to capture hierarchical representations. However, CNN-based models are limited in modeling long-range dependencies due to their inherently local receptive fields imposed by fixed-size convolutional kernels. More recently, Transformers [2], originally developed for language tasks, have been adapted to computer vision in the form of ViT [3], offering an alternative to CNN-based architectures. By leveraging self-attention mechanisms, the ViT model captures long-range spatial relationships and captures global context across entire images. ViT divides images into fixed-size patches, which may result in coarse, cubical segmentations along object boundaries, limiting precise boundary localization of target regions.

[*] mkhalaf@uwaterloo.ca

In this paper, we propose an explainable medical image segmentation framework that provides visual explanations in the form of heatmaps. We implement an alternative architecture that combines the strengths of CNNs and transformers. Specifically, we propose a dual-encoder architecture featuring a U-Net variant encoder to preserve original features and capture hierarchical representations, alongside a ViT–based encoder to capture global context across the entire image. The two encoders operate in parallel and are subsequently fused using an attention-gated fusion mechanism, enabling more discriminative feature integration while simultaneously generating model-specific visual explanations. These explainability heatmaps highlight salient regions that contribute most to the segmentation delineation, providing interpretable visual cues for clinical decision-making. Similar to the hybrid architecture of TransU-Net [4], we fuse encoder output in the decoder; however, our integration of CNN-based attention mechanisms enables more discriminative feature fusion and generates model-specific visual explanations. The resulting heatmaps highlight the most salient regions that contribute to the segmentation delineation, providing medical professionals with interpretable visual cues.

Our approach is designed to assist clinicians in identifying anomalies from various kinds of medical images, including ultrasounds or MRIs. In this paper, we examine the case of assisting in identifying possible tumors when processing images for colorectal cancer (CRC) patients. The importance of enlisting AI to enable improved decision-making in this context is underscored by current research showing that CRC is the third most common cancer worldwide and the second leading cause of cancer-related deaths [5]. CRC cases often develop from adenomatous polyps, which can be effectively treated if identified and removed at an early stage [6]. We make use of the Kvasir-SEG dataset [7], which contains polyp images with pixel-precise ground truth annotations, capturing the variability in polyp appearance, size, and location encountered in clinical practice. For this task, our framework segments polyps and provides visual explanations that describe polyp characteristics such as size, morphology, and surface features. These explanations can assist endoscopists in assessing adenoma characteristics and potential malignancy risk, supporting more informed decisions about polyp removal and follow-up protocols.

Our primary contributions are: (1) a dual encoder architecture for accurate image segmentation, (2) attention-based fusion to combine independent encoders, and (3) model-specific visual explanations that are faithful to the segmentation decision-making process [1].Our aim is to enhance clinical workflows through interpretable AI-assisted diagnosis. We offer results to show the value of our particular approach and return towards the end of the paper to discuss important extensions for future work.

2. Related Works

2.1. Medical Image Segmentation Architectures

U-Net [1] has become the foundational architecture for medical image segmentation. Its encoder-decoder structure with skip connections enables the network to capture both high-level semantic information and fine-grained spatial details, making it effective for segmenting medical images where precise boundary delineation is crucial. The symmetric design allows for propagation of features from the contracting path to the expansive path, facilitating accurate localization. The U-Net architecture has spawned numerous variants tailored to specific medical imaging challenges. However, one of the major drawbacks is the skip connection that indiscriminately concatenates features.

[1]Code is available at https://github.com/MaudDK/MedSeg-XAI-AGFusion

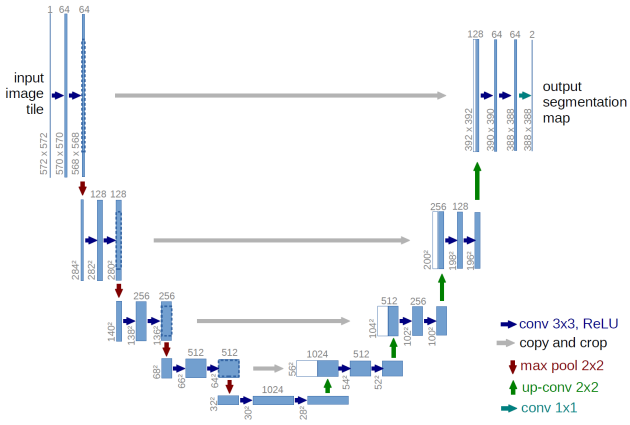

Figure 1. U-Net architecture [1]

U-Net adopts an encoder–decoder architecture with a distinctive U-shaped structure, where skip connections link matching stages in the encoder and decoder. In the encoder, each level applies a 3×3 convolution followed by batch normalization and a ReLU (Rectified Linear Unit) activation, and then uses a 2×2 max-pooling layer with stride 2 to downsample the feature maps. As the spatial resolution decreases, the number of feature channels is typically doubled at each stage. In the decoder, the network progressively upsamples the representations to restore spatial detail while halving the number of channels at each level. After every upsampling operation, two successive 3×3 convolutions are applied, and this process continues until the feature maps return to the original input resolution. Figure 1 depicts the visual representation of the U-Net.

ResUNet [8] integrates residual learning principles from ResNet into the U-Net architecture to address the vanishing gradient problem and enable training of deeper networks. Incorporating residual blocks in both encoder and decoder paths allows direct gradient flow through skip connections. This design improves feature propagation and enables the model to capture more complex hierarchical representations. The residual connections also help preserve fine-grained spatial information throughout the network. ResUNet has demonstrated improved performance over standard U-Net on various medical imaging datasets, especially for challenging segmentation tasks with complex anatomical structures.

Attention U-Net [9] extends the original U-Net architecture by incorporating attention gates that automatically learn to focus on salient regions of the input image. These attention mechanisms suppress irrelevant features while highlighting important spatial locations, improving segmentation accuracy for smaller targets. The attention gates are integrated into the skip connections, allowing the model to adaptively weight feature maps during the decoding process. This selective focus has proven especially valuable in medical imaging where target structures may vary significantly in size and appearance.

Duck-Net [10] introduces a novel architecture designed for efficient medical image segmentation with reduced computational overhead. The model employs a unique block design that balances receptive field expansion with parameter efficiency, making it suitable for deployment in resource-constrained clinical environments. Duck-Net demonstrates that careful architectural design can achieve competitive segmentation performance while maintaining computational efficiency.

2.2. **Transformer-Based Medical Segmentation**

TransU-Net [4] represents a significant milestone in medical image segmentation by combining the strengths of transformers and convolutional neural networks. The architecture employs a standard Transformer as an encoder by first processing the input image through multiple convolutional layers to extract features, which are then projected into embeddings compatible with the Transformer input. These initial convolutional features serve as skip connections to the U-Net decoder for precise localization. TransU-Net demonstrated that hybrid CNN-transformer architectures can model long-range dependencies in medical images, which CNN-only methods may struggle to capture. However, TransU-Net's reliance on convolutional preprocessing limits its ability to leverage pure transformer-based feature learning from the outset. In contrast, our work employs a ViT (specifically Swin Transformer) [11] that directly processes patchified images without initial convolution stages, extracting skip connections from deeper hierarchical transformer layers rather than from preliminary CNN features. Additionally, we introduce attention-based fusion mechanisms between dual encoders and focus explicitly on generating model-specific explainability through our hybrid architecture.

Vision Transformer(ViT) [3] advances the original Transformer architecture by adapting it to process images. Originally designed for classification, the ViT divides an input image into patches (typically 16×16 pixels), which are then projected into embeddings compatible with the original Transformer block. This approach allows the architecture to capture long-range dependencies using self-attention mechanisms and model global contexts effectively. While the original ViT performs image classification, in this paper we adapt ViT to segmentation by replacing its classification head with a U-Net-style decoder, enabling it to generate dense segmentation masks.

MedSAM [12] adapts the Segment Anything Model (SAM) for medical imaging by fine-tuning its ViT-based architecture on a large-scale medical image dataset. As a foundation model, MedSAM enables zero-shot and few-shot segmentation across diverse imaging modalities and anatomical structures through prompt-based interaction. In contrast, our framework produces segmentation outputs through an end-to-end architecture with inherent visual explainability, without requiring external prompting mechanisms.

3. **Methodology**

Medical image segmentation traditionally relied on CNNs, with U-Net as the foundational architecture with skip connections. However, CNNs are inherently limited by their local receptive fields, which are constrained by the kernel size. This locality enables CNNs to capture fine-grained spatial details; however, it limits the ability to model long-range dependencies and global contexts across non-adjacent patches within an image. In contrast, ViT leverages self-attention mechanisms that allow each patch to attend to all other patches in the image, enabling the capture of global context. The self-attention computes relationships between all patches, while the positional encodings preserve the spatial information, allowing the model to understand both the semantic content and the arrangement of each image patch. The global context is important in medical imaging, where understanding relationships between distant anatomical structures is essential.

From a modeling perspective, the importance of global context depends on the spatial distribution and complexity of the segmentation targets. For tasks such as left ventricle[2] segmentation in cardiac imaging, the target is a singular contiguous anatomical structure with a consistent location and shape. The local receptive field of CNNs is sufficient for

[2]We explored medical applications other than CRC and polyps as part of our research and elaborate on this topic in the Discussion section

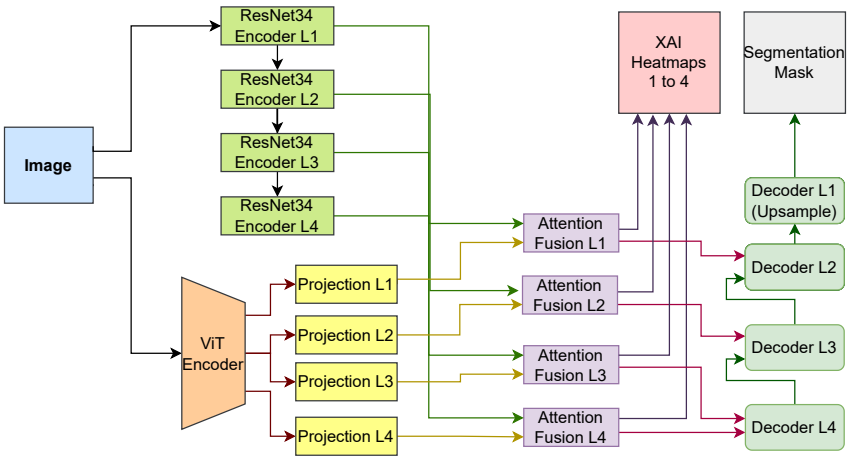

Figure 2. Dual encoder architecture fusing ResNet-34 local features with Swin Transformer global context via learned attention gates at multiple scales.

accurate segmentation in such cases. In contrast, segmentation tasks such as tumor detection may involve multiple diffuse targets that vary in size, appearance, and spatial location. The target distribution across the images requires the model to integrate information over a much larger spatial context. As a result, incorporating global contextual information that ViT offers can be beneficial for capturing relationships between distant yet related regions. Despite the advantages, ViT presents significant computational challenges. Training a ViT from scratch requires demanding computational resources, large datasets, and substantial training time due to the quadratic complexity of self-attention operations. For many healthcare organizations, the training computational requirements can be deterring, limiting their integration and development into clinical workflows.

3.1. **Dual encoder architecture with attention-based fusion:**

Our approach addresses these challenges by combining the strengths of CNNs and ViT through a dual encoder architecture with learned attention-based fusion, shown in Figure 2. We employ a pretrained ResNet-34 CNN encoder paired with a pretrained Swin Transformer (specifically, `swin_base_patch4_window12_384`) as our ViT encoder [11]. By leveraging pretrained weights, we significantly reduce computational requirements and enable the model to perform autonomously without the prompting requirements that limited models like MedSAM.

Our architecture processes input images at 384×384 resolution. The ResNet-34 encoder produces feature maps at multiple scales: 96×96 (128 channels), 48×48 (128 channels), 24×24 (256 channels), and 12×12 (512 channels). The Swin Transformer encoder produces features at corresponding resolutions with higher channel dimensions: 96×96 (128 channels), 48×48 (256 channels), 24×24 (512 channels), and 12×12 (1024 channels). To enable feature fusion, we apply 1×1 convolutional projection layers to reduce the Swin features to match the ResNet channel dimensions at each level.

We fuse features from both encoders at each skip connection level through a modified attention gating mechanism [9]. The mechanism learns to selectively combine the CNN and ViT features based on their relevance to the segmentation task. Specifically, the attention gate takes as input the ResNet skip connection features x and Swin Transformer features g, and computes:

$$\psi = \sigma\left(\text{Conv}_{1\times1}\left(\text{ReLU}\left(\text{BN}\left(\text{Conv}_{1\times1}(g)\right) + \text{BN}\left(\text{Conv}_{1\times1}(x)\right)\right)\right)\right) \tag{3.1}$$

$$\text{fused} = \psi \cdot x + (1 - \psi) \cdot g \tag{3.2}$$

where σ denotes the sigmoid activation, and BN represents batch normalization. The learned attention map $\psi \in [0, 1]$ determines the contribution of each encoder at every spatial location. When attention values approach 1, the fused features rely more heavily on the ResNet's local details; when they approach 0, the Swin encoder's global context dominates. This weighting is learned end-to-end during training and varies spatially across the image, allowing the model to adaptively leverage global and local information where needed.

We use the Swin encoder features as the gating signal (g) because its global understanding of the image provides high-level semantic guidance to determine which local CNN features are most relevant. This aligns with the principle that global context should inform local feature selection. The Swin encoder's holistic view of anatomical structures can identify which fine-grained details captured by the ResNet warrant emphasis in the final segmentation.

3.2. **Interpretable attention maps:**

An essential advantage of the attention-based fusion mechanism is that it generates model-intrinsic visual explanations. The attention maps produced at each encoder level represent the model's actual decision-making process, showing which encoder features contributed to the segmentation at different spatial locations. This fundamentally differs from model-agnostic post-hoc explanation methods such as Grad-CAM [13], which compute gradient-based saliency maps after inference and may not faithfully represent the true reasoning of the model. Our interpretable attention maps are generated during forward propagation as part of the fusion process, providing a more faithful explanation of how the model arrived at its segmentation. Since attention fusion occurs at multiple skip connection levels, our framework generates attention maps at different scales (12×12, 24×24, 48×48, and 96×96). Currently, we utilize individual attention maps from each decoder level for visualization. However, combining these multi-scale attention maps could provide a more comprehensive explanation.

3.3. **Decoder Architecture:**

The decoder follows a standard U-Net structure with skip connections. At each decoder block, we upsample the features by a factor of 2 using bilinear interpolation, concatenate with the corresponding fused features from the skip connections, and apply two 3×3 convolutional layers with batch normalization and ReLU activation. The decoder progressively upsamples from 12×12 to 96×96 resolution. A final upsampling module increases the resolution to the original 384×384 input size, followed by a 1×1 convolution to produce the segmentation mask.

4. **Results & Experiments**

We employ various training strategies to determine the most effective approach for pretraining the models. Training ViT of various sizes, ResNet-34, and with U-Net decoders from scratch on an unmodified dataset demonstrates that ViT struggles to generalize and perform accurate segmentation, while the ResNet-34 model adapts quickly. Further experimentation with augmentation strategies vastly boosts model metrics, indicating that ViT requires more variability in the dataset to generalize effectively. Finally, using pretrained weights on all models and fine-tuning on the augmented dataset yields the greatest performance increase across all models, with ViT models performing similarly to CNN-based

models. We conclude that for ViT models to achieve metrics comparable to CNN-based encoders, significantly more pretraining is required; however, even when metrics are similar, CNN-based encoders remain more effective in training time, inference time, and data efficiency. Table 1 presents a comprehensive comparison of these training strategies.

Table 1. Performance comparison across different training strategies

Model	From Scratch				Augmented				Aug & Pretrained			
	Dice	**IoU**	**Prec.**	**Rec.**	**Dice**	**IoU**	**Prec.**	**Rec.**	**Dice**	**IoU**	**Prec.**	**Rec.**
ViT-Tiny	0.603	0.480	0.718	0.636	0.724	0.618	0.787	0.782	**0.872**	**0.809**	**0.869**	**0.915**
ViT-Small	0.645	0.527	0.709	0.709	0.697	0.592	0.743	0.759	**0.880**	**0.810**	**0.901**	**0.899**
ViT-Base	0.607	0.480	0.670	0.679	0.674	0.558	0.669	0.801	**0.875**	**0.807**	**0.890**	**0.905**
ResNet-34	0.854	0.788	0.890	0.874	0.882	0.817	0.894	0.910	**0.893**	**0.837**	**0.910**	**0.910**

Moving forward with data augmentation combined with pretrained weights, we investigate whether skip connections are essential to our dual encoder design. Skip connections are hypothesized to facilitate rich feature propagation from encoder to decoder, preserving fine-grained spatial information. As shown in Table 2, skip connections consistently improve segmentation performance across all encoders, with Dice improvements ranging from 0.9% to 1.1%, validating their importance in preserving spatial information and enhancing feature propagation.

Table 2. Performance comparison of pretrained models with and without skip connections

Model	Skip Connections	Dice	IoU	Precision	Recall
ResNet-34 UNet	With Skip	**0.893**	**0.837**	0.910	**0.910**
	No Skip	0.883	0.819	**0.925**	0.886
DeepLabV3+	With Skip	**0.890**	**0.825**	**0.890**	0.923
	No Skip	0.880	0.814	0.881	**0.923**
Swin Transformer	With Skip	**0.899**	**0.840**	0.897	**0.935**
	No Skip	0.890	0.829	**0.901**	0.920

Based on the ablation results, we select the Swin Transformer as our transformer-based encoder due to its superior performance (0.899 Dice score [3]) and hierarchical feature representation capabilities, and ResNet-34 for its better inference times and comparable metrics (0.893 Dice score). Consequently, our dual encoder architecture combines the pretrained Swin Transformer and ResNet-34, leveraging global context modeling and efficient local feature extraction.

We then conduct experiments to determine the optimal attention guidance strategy and parameter freezing configuration (Table 3). Incorporating attention mechanisms provides consistent improvements over the no-attention baseline, with Swin-guided attention achieving the highest Dice score (0.907). Single attention guidance (Swin-guided) outperforms dual-guided configurations, suggesting selective feature emphasis from the transformer encoder is more effective than bidirectional attention fusion. Parameter freezing in dual-guided architectures maintains performance while reducing computational overhead. The Swin-guided Attention Dual Encoder achieves the highest Dice (0.907) and IoU[4] (0.855), demonstrating that strategic attention guidance from the transformer pathway enhances segmentation accuracy.

[3]The Dice score is used to indicate the overlap or similarity between samples. It ranges from 0 to 1, with 1 indicating perfect agreement.

[4]IoU is a key computer vision metric that measures the accuracy of an object detector by quantifying the overlap between a predicted bounding box and the ground truth box. It calculates the ratio of the intersection area to the union area of these boxes, ranging from 0 (no overlap) to 1 (perfect match).

Table 3. Performance comparison of Attention Guidance & Parameter Freezing in dual encoder architectures

Model	Attention Guidance	Frozen	Dice	IoU	Precision	Recall
No Attention Dual Encoder	None	Unfrozen	0.902	0.844	0.916	0.922
Attention Dual Encoder	Swin-guided	Unfrozen	**0.907**	**0.855**	**0.926**	0.921
	Res34-guided	Unfrozen	0.904	0.850	0.921	0.919
Double Attention Dual Encoder	Dual-guided	Frozen	**0.906**	**0.853**	**0.922**	0.923
	Dual-guided	Unfrozen	0.901	0.846	0.907	**0.934**
Double Weighted Attention Dual Encoder	Dual-guided	Frozen	**0.903**	**0.848**	**0.906**	0.934
	Dual-guided	Unfrozen	0.902	0.841	0.894	**0.941**
Weighted Double Dual Encoder Attention	Dual-guided	Unfrozen	0.900	0.846	0.927	0.909

Table 4. Performance comparison of best models across different architectures

Model	Dice	IoU	Precision	Recall
Dual Encoder Architectures				
Attention Dual Encoder (Swin-guided)	**0.907**	**0.855**	**0.926**	0.921
Transformer Architectures				
Swin Base (Pretrained, Skip)	0.899	0.840	0.897	0.935
ViT-Small (Pretrained)	0.880	0.810	0.901	0.899
ViT-Base (Pretrained)	0.875	0.807	0.890	0.905
ViT-Tiny (Pretrained)	0.872	0.809	0.869	0.915
CNN Architectures				
ResNet-34 UNet (Pretrained)	0.893	0.837	0.910	0.910
DuckNet (Augmented)	0.889	0.824	0.898	0.917
DuckNet-34 (Augmented)	0.881	0.816	0.900	0.905

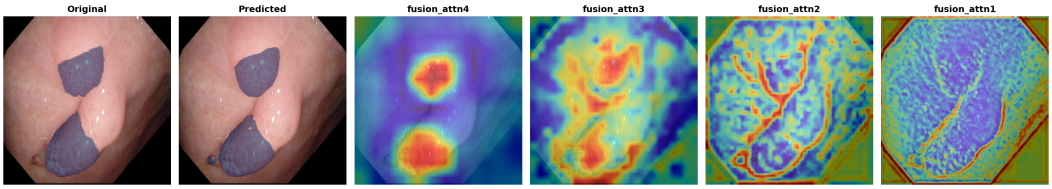

Figure 3. Segmentation and attention visualization.

Table 4 compares our best dual encoder against state-of-the-art transformer and CNN-based segmentation models. The Attention Dual Encoder (Swin-guided) achieves the highest overall performance (Dice 0.907, IoU 0.855), outperforming all individual encoder architectures. Among transformers, Swin Base achieves 0.899 Dice, while ViT variants range from 0.872 to 0.880. Among CNNs, ResNet-34 UNet achieves 0.893 Dice, and DuckNet variants (0.889, 0.881). These results demonstrate that our dual encoder with attention-guided feature fusion effectively combines the complementary strengths of transformer and CNN encoders, achieving superior segmentation performance compared to single-encoder architectures.

Figure 3 attention maps from our architecture reveal a clear hierarchical focus. The first two maps primarily highlight the edges of the polyp, capturing boundary information, while the subsequent maps concentrate on the polyp region itself, guiding the network to generate the final segmentation. This progression demonstrates how the model sequentially refines spatial features from edges to the full object.

5. Discussion and Future Work

This work targets a core barrier to clinical adoption of deep segmentation systems: even when masks are accurate, clinicians still need actionable, human-readable justification for why particular regions were delineated. The proposed framework addresses this gap by pairing (i) a dual-encoder segmentation backbone (CNN + ViT) with (ii) a model-specific explainability heatmap derived from attention fusion. As mentioned earlier, our framework should be applicable to any number of possible medical scenarios, even though we illustrated its significance for the important topic of identifying regions of images where possible tumors exist or do not exist. The heatmaps we produce enable progressively more valuable insights to clinicians. Certainly, a really critical area to explore is one where a sequence of images with motion, over a period of time, needs to be understood. An ideal medical application for this study is that of cardiology [14]. As mentioned earlier, we started to consider this medical concern, interested in determining whether patients at risk of developing cardiac episodes could be discerned from the medical images. Towards the consideration of longer sequences, we can extend the pipeline from single-frame inference to echocardiogram sequences, where consistency over time is clinically essential; recent advances in promptable image/video segmentation (e.g., SAM 2 with streaming memory) motivate evaluation of temporally stable segmentation and explanation behaviors under realistic motion and acquisition variability.

When it comes to providing an explainable AI solution for medical practitioners, a key question is whether explanations actually improve clinician decision-making or sometimes harm it. Panigutti et al. [15] show that explanations in clinical decision support can increase advice-taking, but also reveal some areas where there may be hesitations: clinicians may request explanations, yet report cognitive overload when explanation interfaces become too dense, and concerns about over-reliance or replacement remain salient. This aligns with broader evidence that the benefit of saliency-style explanations is conditional: Muller's systematic review finds that saliency maps can help [16]. Importantly, recent controlled evidence suggests that explainability can improve task outcomes in expert human–AI collaboration: Senoner et al. report measurable gains in expert performance when heatmaps accompany AI outputs compared to black-box assistance [17]. Together, these findings motivate a cautious interpretation: it is not enough to "add a heatmap"; explanations must be tuned to the clinical task, cognitive workload, and decision context, and should be empirically evaluated with end users rather than assumed beneficial by design.

These human factors also highlight why how saliency is produced and presented matters. Alqaraawi et al. found that saliency maps can help users learn model-sensitive features, yet offer limited support for anticipating model outputs on new cases, suggesting that saliency alone may not provide the type of reasoning clinicians need [18]. Zhao et al. further show that graphical perception factors, such as visual encoding choices, alignment, and saliency map properties, substantially shape how people interpret saliency-based explanations [19]. These results are directly relevant to the heatmap component in this work: even if attention fusion maps offer a more architecture-tied signal than model-agnostic CAM variants, their clinical utility will still depend on visualization design (e.g., overlay strategy, calibration, thresholding, and whether the map supports rapid "at-a-glance" verification rather than demanding interpretive effort). All of these considerations continue to motivate continued effort to proceed with some greater engagement of our user base of medical experts, as an important step forward in the future.

We are encouraged by some companion efforts to promote new pathways for explainable AI in medical contexts. The work of Khalaf et al. [20] with its comparative study has made important inroads by demonstrating that model-specific solutions can support attention and thus explanation, in ways that provide accuracy on par with some of the best U-net models that lack this XAI capability. Some kind of deeper engagement with the user base

of clinicians may yet provide the best insights into what medical professionals prefer to experience for their explanations. The kind of qualitative feedback provided in Paniguitti et al. [15], (comments provided by clinicians to accompany any quantitative evaluation) may serve as an inspiring first step. Another avenue for acquiring a critical, deeper understanding of benefits arising from model-specific approaches is to apply them to a wide variety of medical concerns; beyond our examination of tumors for colorectal cancer, we have already begun to study datasets for breast cancer diagnoses, and have mentioned as well above some first steps exploring cardiovascular applications.

There has also been a theme in some current research of critiquing whether popular techniques like GRADCAM used for XAI in the context of computer vision are sufficiently faithful to the decision process, to meet user requirements. The work of Wu et al. [21] and that of Nguyen et al. [22] both develop novel pathways for spatial saliency refinement within that context, considering ways of extending the integration of attention as new food for thought for our continued study. Another thread for future work is to study existing datasets to see whether the way they are coded may pose challenges to effective segmentation. In preliminary experiments of ours on the CAMUS dataset for left ventricular segmentation [23] and with a simpler Attention-VIT architecture, the greyscale nature of the echocardiogram images made it challenging to differentiate between the left ventricle, myocardiam and surrounding cardiac structures. Future work can explore whether certain datasets will be inherently more difficult to use, perhaps leading to efforts to expand the data collection process.

5.1. **Expanding to integrate textual explanations as well**

The larger agenda for our research project is to also provide clinicians with textual explanations to enrich their understanding of the images. We would imagine using an RAG + LLM component that converts retrieved medical content into textual explanations aligned with the detected target. This "visual" + "textual" design would be to support clinical sensemaking beyond what post-hoc saliency alone can provide. We view the RAG-constrained approach to be a valuable complement to what LLMs would produce, after suitable prompting. We may gain insights from Valerio et al., [24] who support LLM-based report generation and are also encouraged by work that emphasizes clinician-facing reporting and interpretability [25]. This work also shows that allowing explanations to reference up-to-date clinical knowledge rather than relying on parametric memory is increasingly important. Recent work on visual medical entity linking (VELCRO) reinforces this point: it shows that querying a knowledge base with the whole image can retrieve overly generic entities, while using only a cropped region-of-interest can lose crucial context; by aligning segmentation-derived contextual RoI embeddings with KB entries via contrastive learning, VELCRO improves linking accuracy and highlights the importance of context-aware grounding for region-specific clinical text [26]. At the same time, evaluation remains a central open issue: De Bona et al. propose using LLMs to replicate aspects of human participation to streamline XAI evaluation when user studies are costly or hard to scale [27]. While such LLM-based evaluation should not replace real clinician studies, it may provide a useful complementary layer (e.g., rapid iteration on explanation readability, consistency, or guideline adherence) before resource-intensive clinical validation.

With more detailed proposals for integrating textual output, we can move beyond "heatmap-only" explanations toward clinically interpretable, concept-grounded evidence. Specifically, the localized medical concepts (e.g., boundary ambiguity, artifact cues, morphology descriptors) can be attached to regions emphasized by the fusion mechanism, enabling explanations that reflect the vocabulary clinicians use when judging segmentation reliability. This will provide two concrete technical benefits: (i) explanations that are more testable (concept

agreement can be measured) and (ii) improved model auditing (concept distributions can reveal spurious correlations and dataset biases). Finally, because real-world adoption of generated clinical text depends on expert assessment and collaboration workflows, we will adopt evaluation protocols that go beyond automated text metrics, inspired by recent large-scale radiologist evaluations and assistive editing setups in report generation systems [28].

6. Conclusion

This paper presents an explainable medical image segmentation framework that unifies accurate segmentation with clinician-oriented interpretability by combining a dual-encoder architecture (CNN for fine spatial detail and a transformer for global context) with an attention-based fusion mechanism that yields model-derived visual evidence. The results demonstrate the feasibility of producing segmentation outputs together with complementary visual rationales, while highlighting the broader implication that medical XAI should be treated as a workflow-facing interface: explanations must be faithful, comprehensible, and actionable to support trust calibration and responsible clinical adoption. Our work offers the following advances: a) improved image segmentation accuracy for medical images b) supporting semantic interpretations from ViT in order to enable valued explanations c) producing heatmaps that tell clinicians where tumors have been identified in images that improve upon former solutions d) discussing limitations that may arise to enable researchers to know what kinds of datasets may need augmentation e) producing new advances for explainable AI through an architecture that may support integration of text in addition to visual clarifications.

References

[1] O. Ronneberger, P. Fischer, and T. Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. 2021.

[4] J. Chen et al. "TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers". In: *Medical Image Analysis* 97 (2024), p. 103280.

[5] R. Siegel, K. Miller, H. Fuchs, and A Jemal. "Cancer statistics, 2022." In: *CA: a Cancer Journal for Clinicians* 72.1 (2022), pp. 7–33.

[6] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, and H. D. Johansen. "Resunet++: An advanced architecture for medical image segmentation". In: *2019 IEEE international symposium on multimedia (ISM)*. IEEE. 2019, pp. 225–2255.

[7] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. De Lange, D. Johansen, and H. D. Johansen. "Kvasir-seg: A segmented polyp dataset". In: *International conference on multimedia modeling*. Springer. 2019, pp. 451–462.

[8] Z. Zhang, Q. Liu, and Y. Wang. "Road extraction by deep residual u-net". In: *IEEE Geoscience and Remote Sensing Letters* 15.5 (2018), pp. 749–753.

[9] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al. "Attention u-net: Learning where to look for the pancreas". In: *arXiv preprint arXiv:1804.03999* (2018).

[10] R.-G. Dumitru, D. Peteleaza, and C. Craciun. "Using DUCK-Net for polyp image segmentation". In: *Scientific reports* 13.1 (2023), p. 9803.

[11] 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: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.

[12] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang. "Segment Anything in Medical Images". In: *Nature Communications* 15 (2024), p. 654.

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[14] S. Islam, G. Rjoub, et al. "Machine learning innovations in CPR: a comprehensive survey on enhanced resuscitation techniques". In: *Artificial Intelligence Review* 58.8 (2025), p. 233.

[15] C. Panigutti, A. Beretta, F. Giannotti, and D. Pedreschi. "Understanding the Impact of Explanations on Advice-Taking: A User Study for AI-Based Clinical Decision Support Systems". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 2022.

[16] R. Müller. "How Explainable AI Affects Human Performance: A Systematic Review of the Behavioural Consequences of Saliency Maps". In: *International Journal of Human–Computer Interaction* (2024).

[17] J. Senoner et al. "Explainable AI improves task performance in human–AI collaboration". In: *Scientific Reports* 14 (2024), p. 31150.

[18] A. Alqaraawi et al. "Evaluating saliency map explanations for convolutional neural networks: a user study". In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 2020.

[19] Y. Zhao, M. Li, and M. Berger. "Graphical Perception of Saliency-based Model Explanations". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 2023.

[20] M. Khalaf, R. Cohen, P. Alencar, J. Bentahar, and S. Islam. "A Little Attention: Model-Specific Explainable Medical Imaging". In: *Proceedings of the IJCAI 2025 Workshop on Large Language Models and Generative AI for Health Informatics*. 2025, pp. 1–9.

[21] J. Wu et al. "On the Faithfulness of Vision Transformer Explanations". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

[22] T. T. H. Nguyen, V. T. K. Nguyen, V. B. Truong, Q. K. Nguyen, P. T. L. Nguyen, F. Palma, and Q. H. Cao. "XGD: Explainable AI-Guided Knowledge Distillation with Feature Refinement for Semantic Segmentation". In: *38th Canadian Conference on Artificial Intelligence, Canadian AI 2025, Calgary, AB, Canada, May 26-29, 2025, Proceedings*. Canadian Artificial Intelligence Association, 2025.

[23] S. Leclerc, E. Smistad, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, C. Lartizien, L. Løvstakken, and O. Bernard. "Deep Learning Segmentation in 2D Echocardiography Using the CAMUS Dataset: Automatic Assessment of the Anatomical Shape Validity". In: *International Conference on Medical Imaging with Deep Learning (MIDL 2019), Extended Abstract Track*. HAL Id: hal-02395245, version 1. London, United Kingdom, July 2019.

[24] A. G. Valerio et al. "From segmentation to explanation: Generating textual reports from MRI with LLMs". In: *Computer Methods and Programs in Biomedicine* (2025).

[25] R. Tanno et al. "Collaboration between clinicians and vision–language models in radiology report generation". In: *Nature Medicine* (2025).

[26] K. Carbone, L. Hebert, R. Cohen, and L. Golab. "Visual Medical Entity Linking with VELCRO". In: *Proceedings of Machine Learning Research (ML4H)*. Vol. 297. 2025.

[27] F. B. De Bona et al. *Evaluating Explanations Through LLMs*. arXiv:2410.17781. 2024.

[28] R. Yang, Y. Ning, E. Keppo, et al. "Retrieval-augmented generation for generative artificial intelligence in health care". In: *npj Health Systems* 2 (2025), p. 2.