LViTES: Leveraging vision and text for enhancing segmentation of endoscopic images

Thang La[1], Minh-Hanh Tran[1], Viet-Hang Dao[2,3], and Thanh-Hai Tran[1]

[1] School of Electrical and Electronic Engineering
Hanoi University of Science and Technology
[2] Hanoi Medical University Hospital, Hanoi, Vietnam
[3] Institute of Gastroenterology and Hepatology, Hanoi, Vietnam
hai.tranthithanh1@hust.edu.vn

Abstract. Automatic lesion segmentation in endoscopic images is crucial for mitigating the risk of omissions during analysis, particularly for inexperienced physicians or in situations of medical overload. Traditional segmentation models predominantly rely on pixel-level labeled images, often neglecting auxiliary information such as physicians' diagnostic conclusions. This study proposes a novel approach to harness available lesion information—including segmentation regions, physician conclusions, and supplementary disease descriptions—to improve segmentation efficacy. Our method builds upon the successful integration of CNN and Vision Transformer architectures from the LViT model, originally designed for lung cancer lesion segmentation from X-ray images using dual inputs: images and text. We propose a new framework, namely called LViTES with four key advancements: 1) optimizing the LViT architecture to enhance image feature extraction by incorporating the EfficientNet backbone and integrating Cross-Attention, while also reducing model complexity and parameters; 2) addressing the scarcity of textual descriptions in current datasets by developing a module that generates text from segmentation masks based on attributes like shape, location, size, and quantity; 3) incorporating both image and text inputs during training while allowing adaptive prediction with only image inputs to align with typical use cases; and 4) evaluating model performance using both generated text and physician-provided descriptions. The effectiveness of our approach is validated on three types of lesions—gastric cancer, esophageal cancer (our self-collected datasets), and polyps (Kvasir-SEG dataset)—demonstrating superior performance compared to state-of-the-art methods.

Keywords: Segmentation · Endoscopic images · Deep learning · Transformer · CNN · LLM

1 Introduction

Digestive diseases are a significant global health concern, with millions affected worldwide [22]. The rising prevalence of these diseases underscores the urgent

need for advanced diagnostic tools that can aid in early detection and accurate diagnosis. One promising solution is the automatic analysis and diagnosis of gastrointestinal conditions using artificial intelligence (AI).

Current methods for gastrointestinal endoscopy image segmentation predominantly rely on state-of-the-art convolutional neural networks (CNNs) or Transformers [21], [19], [20]. These techniques leverage the power of CNNs to accurately segment images, which is crucial for identifying abnormalities in the digestive tract. However, recent advancements in large language models (LLMs) have shown the potential to enhance image analysis by integrating language, thereby improving overall diagnostic accuracy [27].

Several models that combine language and image processing have been successfully applied to natural images. Notable examples include CLIP [16] and SAM [6], which have demonstrated significant improvements in various image analysis tasks. Despite their success in natural image domains, there has been limited research on applying these language-image models to medical imaging, particularly in the field of gastrointestinal endoscopy.

A few studies have explored the application of language and image models to chest X-rays [13], but there is currently no model specifically designed for endoscopic image analysis. The primary challenges include the scarcity of large databases and the lack of accompanying textual data. Addressing these challenges could pave the way for significant advancements in medical imaging analysis.

This paper presents an initial experimental study on the integration of language and images to enhance the segmentation quality of medical endoscopic images. We propose a novel approach LViTES (**L**everaging **Vi**sion and **T**ext for **E**ndoscopic **S**egmentation) that utilizes dual inputs: images and text. The text can be derived from doctors' conclusions or automatically generated from the ground truth masks. These modalities are processed through a CNN - Transformer (e.g. LViT) network architecture to extract and cross-interact features, ultimately improving segmentation outcomes.

Our main contributions are as follows: i) an original research combining text and images for gastrointestinal endoscopy image segmentation; ii) an integration of various textual generation methods to address the scarcity of medical documents; iii) a mechanism allowing training with both text and images, while enabling inference using only image inputs. Experimental evaluation across different datasets (gastric cancer, esophageal cancer, and polyps), demonstrated remarkable results. This pioneering study aims to open new avenues for the application of AI in medical imaging, ultimately contributing to better diagnostic tools and improved patient outcomes in the realm of digestive diseases.

2 Related works

2.1 Semantic segmentation of medical images

Semantic segmentation is a method in the field of computer vision and image processing that aims to classify each pixel in an image into different classes or labels.

In this task, the Fully Convolutional Network (FCN) [10] is considered the first end-to-end pixel-to-pixel network to be published. Following this, models such as DeepLab [1] with three main components: Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs have achieved impressive results in medical image segmentation and satellite image processing. Additionally, Mrdff [24, 25] for CT whole heart segmentation were developed based on and improved upon the FCN. Among these, a pioneering model in the medical field is U-Net [17]. Subsequently, various U-Net variants have been developed to enhance performance and accuracy in different medical applications. Notably, UNet++ [28] utilizes more complex skip connections between the encoder and decoder layers to improve detail and accuracy in segmentation. Attention U-Net [15] incorporates attention mechanisms into the architecture, and DenseUNet [7] is a variant combining U-Net and DenseNet. However, most methods face limitations with the amount of data and heavily depend on data quality, resulting in constrained model performance in generalization. Some recent approaches [8, 11, 12, 23] employing semi-supervised learning methods have partially addressed the issue of data and label scarcity.

2.2 Text and image combination for image analysis

In recent years, vision-language models have gained importance and become a focal point of development. Among these, CLIP [16] stands as the first pioneering vision-language pre-trained (VLP) model, yielding impressive results in vision-and-language tasks and improving accuracy over previous pre-trained models based solely on images or text [18]. CLIP employs a neural network incorporating transformers, specifically utilizing the Vision Transformer (ViT) for image processing and a BERT-like transformer variant for text processing. The model is trained on a large dataset of images and text from the internet, combining image data with textual descriptions. Despite the advantages CLIP brings, it still faces challenges such as handling specific problem domains, computational efficiency, and model complexity.

Recent research has further explored the integration of image-text information to enhance performance in image segmentation tasks, with models like VLT [4], LAVT [26], and TransVG++ [2]. VLT can generate queries from textual descriptions to guide the model during image segmentation. Conversely, LAVT employs early fusion mechanisms and pixel-word attention to optimize image segmentation based on linguistic guidance. TransVG++ uses a Language Conditioned ViT trained to embed textual description information into the image processing workflow, leveraging attention mechanisms to synchronize information from images and text, thus enhancing object recognition accuracy.

However, these models are not specifically designed for medical images. Natural images and medical images exhibit numerous differences. Medical images of different body parts also have distinct characteristics, often lacking clear delineation and uniformity in size and shape. Consequently, directly applying computer vision models trained on natural images to medical image analysis poses

challenges. Furthermore, textual descriptions of medical images differ significantly from those of natural images, resulting in weak correlations between text and images in the medical context.

To address these challenges, the LViT [9] model was developed, designed specifically for the medical field. It employs a hybrid CNN-Transformer architecture to retain both local and global features, utilizes only the Embedding layer to transform text features, reducing parameters and computational cost, and incorporates LV (LanguageVision) loss to supervise training of unlabeled images using direct text information. LViT was validated on chest X-ray images. Combining of language and vision for endoscopic images are underexplored.

3 Proposed method

3.1 General framework

In this paper, we reply upon LViT for endoscopic image segmentation with the training from both images and textual description. However, we have made several significant improvements based on the successful combination of CNN and Transformer in LViT, while also using the EfficientNet Backbone to image feature extraction. Additionally, a cross-attention mechanism has been added during the decoder stage to facilitate closer interaction between the features, thereby improving segmentation performance and synchronizing information between the different data streams. This combination not only enhances the model's analytical capabilities but also improves its ability to effectively handle lesion-related descriptions. Our proposed framework for lesion segmentation is illustrated in Fig. 1.

- At the training phase, it takes two inputs (the pair of original image and the corresponding mask annotated by the experts, and a textual description of the lesion). The textual description can come from two different sources: the medical report produced by doctors or automatically generated by our algorithm from the groundtruth mask. The textual stream goes through a text embedding to provide a text representation while the image is fed into the EfficientNet Backbone for feature extraction. Text features and image features at different layers interact in an attention model.
- At the inference phase, the model can take an image with or without a text description to generate a segmentation result.

In the following, we will detail each step of the framework.

3.2 Text-based description processing

Our disease description texts are generated from two sources: medical reports provided by the hospital and auto-generated texts from masks, as shown in Fig. 1.

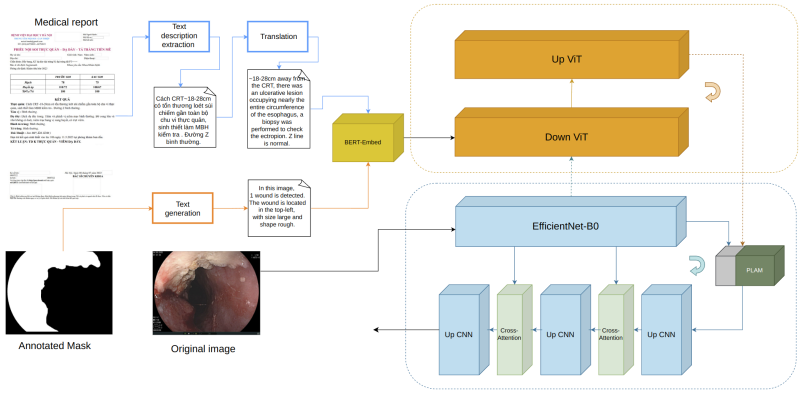

Fig. 1: Proposed framework LViTES for endoscopic image segmentation.

Text extraction As we mentioned, there is no dataset of endoscopic images which contains both binary masks and the corresponding textual description about the lesion. To test our method, we have collected a dataset which contain two lesion categories: gastric cancer, esophageal cancer. The textual descriptions of the lesion are stored in structured pdf format. We develop a script to acronymize the patients identity. We then extract the conclusion about the lesion part. The top-left of Fig. 1 illustrates a medical report for a given patient. We observe that during an endoscopic examination, the doctor reports all relevant issues of the digestive system, such as those involving the esophagus, stomach, duodenal bulb, duodenum, etc. However, if a lesion is identified as Gastric Cancer or Esophageal Cancer, we focus exclusively on extracting information related to the that lesion.

Text translation As the collected texts are in Vietnamese or may be any other languages, therefore, we translated them into English for text encoding. In our work, our self-collected dataset contain textual description in Vietnamese. We use the VinAI Translate model [14] to translate text in Vietnamese into English. In Fig. 1, we illustrate an example where the Vietnamese text about Esophageal cancer is extracted as "Cách CRT~18-28cm có tổn thương loét sùi chiếm gần toàn bộ chu vi thực quản, sinh thiết làm MBH kiểm tra. Đường Z bình thường." It is translated into English as "~18-28cm away from the CRT, there was an ulcerative lesion occupying nearly the entire circumference of the esophagus, a biopsy was performed to check the ectropion. Zline is normal."

Mask-based text description generation The textual description of the lesion is not always available. For example, for KVasir-SEG dataset, only groundtruth mask is provided. To address the lack of textual description, which is a common problem in practical situation, we propose a novel approach to generate text

descriptions from the image data itself. This approach mitigates the dependency on textual annotations provided by physicians, making it particularly useful in situations where such annotations are scarce or unavailable.

Our proposed method includes following steps. First, we detect contours from the mask image using Canny. Then we apply connected component analysis to determine the legion regions in terms of number, location, and shape. For location, we divide the image into different regions (center, top-middle, bottom-middle, left-middle, right-middle, top-left, top-right, bottom-left, bottom-right, left, right, top, bottom) and determine the location of each wound within these regions. For shape description, we calculate characteristics of the wounds such as area, perimeter, roundness, shape, and size. Finally, based on the number, location, and characteristics of the wounds, we create a natural language description of the image. Fig. 2 illustrated the main steps for textual description generation from the ground truth image.

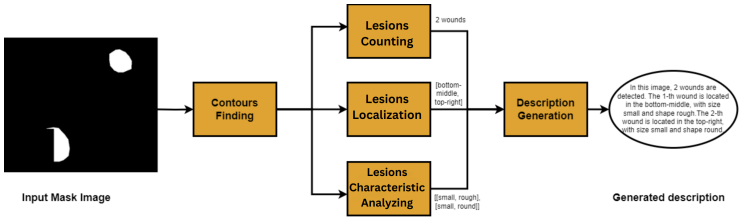

Fig. 2: Main steps for textual description generation from the groundtruth image.

3.3 Model architecture

The input of our model in the training phase composes of two components: an original image and its groundtruth and the corresponding textual description (real of generated). Text embedding is computed using the well-know pretrained BERT [3] while image pass through an EfficientNet-B0 for extracting visual features. To capture the correlation between visual and textual features, both $\mathbf{T_{embed}}$ and $\mathbf{I_{embed}}$ go through a down ViT then an up ViT. The output $\mathbf{Z}_{reconstruct}$ from the up ViT and the visual feature continue to pass through a Pixel Level Attention Module (PLAM) before entering to a up sampling module for generating the segmentation map.

After the image and text features have been extracted through their respective branches, they pass through double Vision Transformer (ViT) blocks to optimize and combine the information before entering the decoder stage. The Vision Transformer block is completely inspired from the LViT model [9]. However, it is noted that our architecture LViTES makes some differences. First, for the image embedding, LViT utilized fours down CNNs which contains Conv, BatchNorm(BN), and ReLU activation layers for image embedding. Instead, LViTES

employed EfficientNet-B0 as explained in the previous subsection. Second, we keep only one down ViT and one up ViT in the ViT U-block that capture interaction between text and image embeddings while the original LViT contains four down and four up ViTs. We reduce the number of ViT blocks to avoid overfitting and enhance the accuracy of the model as presented in our expriments. We detail our architecture LViTES as follows:

– **Down ViT Block**:
  - The image embedding $\mathbf{I}_{\text{embed}}$ and text embedding $\mathbf{T}_{\text{embed}}$ are processed simultaneously through Transformer layers to capture relationships between image patches and text features.
  - $\mathbf{I}_{\text{embed}}$ is passed through PatchEmbedding, while $\mathbf{T}_{\text{embed}}$ is aligned using a CTBN block (Conv, BatchNorm, ReLU), and both are merged:

$$\mathbf{x}_{\text{merged}} = \text{PatchEmbedding}(\mathbf{I}_{\text{embed}}) + \text{CTBN}(\mathbf{T}_{\text{embed}})$$

  - The merged feature is passed through a Transformer's Multi-Headed Self-Attention (MHSA) and MLP layers:

$$\mathbf{x}' = \text{MHSA}(\text{LN}(\mathbf{x}_{\text{merged}})) + \mathbf{x}_{\text{merged}}$$

$$\mathbf{Z} = \text{MLP}(\text{LN}(\mathbf{x}')) + \mathbf{x}'$$

  - We briefly write the above calculation process as follows:

$$\mathbf{Z} = f^1_{\text{ViT}}(\mathbf{I}_{\text{embed}}, \mathbf{T}_{\text{embed}})$$

– **Up ViT Block (Reconstruction)**:
  - The output $\mathbf{Z}$ from the first block is reconstructed by the second ViT block. Gives output the same size as the visual feature:

$$\mathbf{Z}_{\text{reconstruct}} = f^2_{\text{ViT}}(\mathbf{Z})$$

  - The second block introduces a CTBN update:

$$\mathbf{x}' = \mathbf{x} + \text{CTBN}(\mathbf{x})$$

– **Pixel-Level Attention Module (PLAM)**:
  - PLAM merges local image features with semantic text features. It uses Global Average Pooling (GAP) and Global Max Pooling (GMP) on the input X. For our model, X represents the sum of $\mathbf{Z}_{reconstruct}$ from the up ViT and the visual feature:

$$\mathbf{z}_{\text{GAP}} = \text{GAP}(\text{RELU}(\text{Conv2D}(X))), \quad \mathbf{z}_{\text{GMP}} = \text{GMP}(\text{RELU}(\text{Conv2D}(X)))$$

  - The combined output:

$$\mathbf{z}_{\text{concat}} = \text{CONCATENATE}(\mathbf{z}_{\text{GAP}}, \mathbf{z}_{\text{GMP}}, \mathbf{z}_{\text{GAP}} + \mathbf{z}_{\text{GMP}})$$

- This is passed through an MLP and scaled:

$$y = X \times \text{MLP}(\mathbf{z}_{\text{concat}})$$

– **Decoder**:
  - $\mathbf{Z}_{\text{reconstruct}}$, the initial image feature $\mathbf{I}_{\text{embed}}$, and PLAM outputs are combined.
  - The decoder applies convolutional layers, upsampling, and cross-attention between image and text features for final output generation.

In this work, we inherit the loss function design from the LViT model, which consists of a combination of **Dice loss** and **Cross-Entropy loss**. The loss function is used to guide the model in learning accurate segmentation by minimizing errors in both labeled and unlabeled data. To further improve the robustness of the model and mitigate overfitting, we incorporate an **L2 regularization** mechanism. The L2 regularization term penalizes large weights by adding the squared magnitude of weights to the loss function. This helps reduce the complexity of the model and prevents it from overfitting to the training data.

4 Experiments

4.1 Datasets and Implementation details

In this study, we utilize two datasets: Kvasir-SEG [5] and IGHEndoLesion-Image-Report, which are divided into training, validation, and test sets in a ratio of 7:1:2, as shown in Table 1.

Kvasir-SEG: This is a widely used dataset for tasks related to the evaluation of models for detecting and segmenting lesions in the form of polyps, comprising 1,000 images with resolutions ranging from 332×487 to 1920×1072. This dataset provides only ground-truth image, not textual description. We must apply text generation module to train the segmentation model.

IGHEndoLesion-Image-Report: This dataset consists of 2,667 pairs of image reports with a resolution of 1280×995, where 1,383 pairs are related to esophageal cancer and 1,284 pairs pertain to gastric cancer. Each patient may have more than one pair of image reports. Each image is accompanied by a detailed report regarding the patient's condition, provided by physicians who directly conducted the examinations. These reports contain critical information about the diagnosis and severity of lesions. Each report includes three main sections of descriptive information: characteristics at the cancer site; detailed descriptions of the lesions such as size, shape, and location; and annotations regarding the patient's condition. Specifically, reports concerning gastric cancer have an average length of 232 characters per report and describe features such as gastric fluid, mucosa, curvature, antrum, and cardia. Reports related

to esophageal cancer have an average length of 218 characters per report and describe features such as veins, esophageal lumen, Z-line (transition zone of the mucosa), and mucosa. The images and medical reports are provided by the Gastroenterology and Hepatology Institute of Hanoi Medical University.

Table 1: Summary of experimented datasets

Dataset	IGHEndoLesion-Image-Report		Kvasir-SEG
Lesion Type	Gastric Cancer	Esophageal Cancer	Polyp
Train set	898	968	700
Val set	128	138	100
Test set	258	277	200
Total	1284	1383	1000

The proposed approach is implemented using PyTorch. The main server parameters are as follows: the operating system is Ubuntu 20.04.6 LTS, the CPU is an Intel(R) Xeon(R) Gold 6130, the GPU is a single NVIDIA RTX 4090, and the memory capacity is 128 GB. The initial learning rate is set to 3×10^{-4} for all datasets. We also use an early stopping mechanism, which halts training if the model performance does not improve after 40 epochs. The default batch size is 16 for both datasets, and all images are resized to 224×224. Additionally, the cosine learning rate schedule is applied during training.

4.2 Experimental results

We evaluated our model on three datasets: Gastric Cancer, Esophageal Cancer, and Kvasir-SEG. For the first two, we compared it to the original LViT model using text generated by Mask-based text description generation (G) and text from medical records (D). For Kvasir-SEG, we compared our results to state-of-the-art methods and conducted additional experiments to assess the effects of not using text (W) and only using text during training (IW). It is noted that for Kvasir-SEG dataset, we generated textual description about the polyps.

Table 2: Experimental results on the Gastric Cancer dataset.

Method	mIoU	DSC	Recall	Precision
LViT-D	0.7200	0.8125	0.8418	0.9050
LViT-G	0.7437	0.8318	0.8754	0.8906
LViTES-D	0.7618	0.8829	0.9114	0.9309
LViTES-G	**0.7855**	**0.8941**	**0.9127**	**0.9449**

Table 3: Experimental results on Esophageal Cancer dataset.

Method	mIoU	DSC	Recall	Precision
LViT-D	0.6707	0.7828	0.8746	0.8835
LViT-G	0.7006	0.8068	0.8725	0.9024
LViTES-D	0.7579	0.8596	0.9198	0.9411
LViTES-G	**0.7634**	**0.8702**	**0.9288**	**0.9497**

Table 4: Experimental results on the Kvasir-SEG dataset.

Method	mIoU	DSC	Recall	Precision
U-Net	0.7472	0.8264	0.8504	0.8703
U-Net++	0.7420	0.8228	0.8437	0.8607
ResU-Net++	0.5341	0.6453	0.6964	0.7080
HarDNet-MSEG	0.7459	0.8260	0.8485	0.8652
ColonSegNet	0.6980	0.7920	0.8193	0.8432
UACANet	0.7692	0.8502	0.8799	0.8706
UNeXt	0.6284	0.7318	0.7840	0.7656
TransNetR	0.8016	0.8706	0.8843	0.9073
LViT-G	0.8040	0.8779	0.9039	0.9172
LViTES-G (Our)	**0.8642**	**0.9306**	**0.9271**	**0.9363**

Table 5: Other experimental results on the Kvasir dataset.

Method	mIoU	DSC	Recall	Precision
LViTES-W	0.8125	0.8994	0.8872	0.8997
LViTES-IW	0.8311	0.9061	0.8988	0.9124
LViTES-G	**0.8642**	**0.9306**	**0.9271**	**0.9363**

Table 2 and Table 3 show the results obtained by our model, LViTES, compared to the original LViT in two scenarios: using textual descriptions provided by doctors and using textual descriptions generated by our modules. It is shown that in both scenarios, our proposed model, LViTES, outperformed LViT. Specifically, the generated textual descriptions provided slightly better performance than those provided by doctors. This can be explained by the fact that when textual descriptions from doctors are unavailable, the generated text becomes highly significant. Table 4 compares our model, LViTES-G, which uses generated textual descriptions, and demonstrates that it outperformed all existing state-of-the-art (SOTA) methods such as U-Net, U-Net++, ResU-Net++, HarDNet-MSEG, ColoSegNet, UACANet, UNext, and TransNetR. LViTES-G achieved a mIoU that is 6.02% higher than the original LViT-G

Table 6: Qualitative results on the Kvasir-SEG dataset.

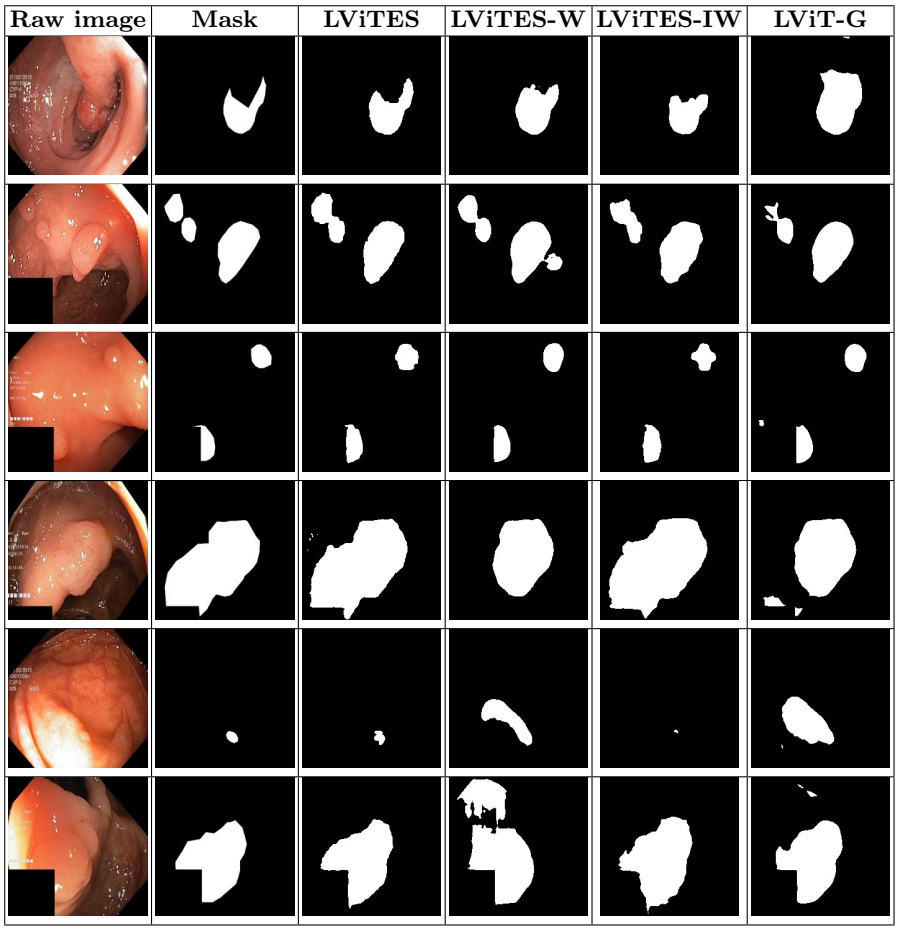

Table 7: Qualitative results on the Gastric Cancer and the Esophageal Cancer datasets.

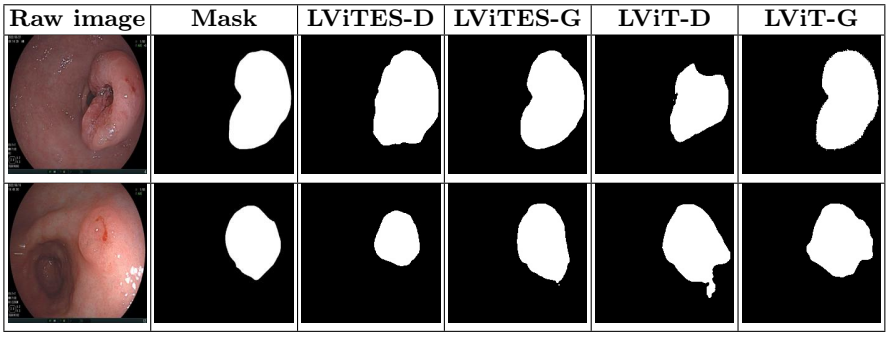

Table 5 presents the results of our ablation study, and shows the effects of modality on segmentation performance. LViTES-W is the model we trained using only images, without text, while LViTES-IW is the model we trained using only text, without images. We compared these with the final model, LViTES-G, which was trained using both text and images. We observed that using both modalities for training increased the mIoU of LViTES-G by 3.31% and 5.17% compared to LViTES-IW and LViTES-W, respectively.

Table 6 visualizes some segmentation results obtained by LViTES on Kvasir-SEG, compared against the original LViT. It is noted that LViTES produces segmentation maps that best fit the ground truth, whereas LViT sometimes produces larger or smaller lesions than the ground truth. Table 7 illustrates examples comparing results using doctors' textual descriptions against generated textual descriptions. Due to the additional information in the doctors' textual descriptions, the segmentation results are sometimes more biased. In contrast, the generated descriptions are based on ground truth images, helping to focus more on the region of interest by describing the shape, location, and number of lesions. As a result, the segmentation regions are more accurate.

5 Conclusion

In this paper, we presented a novel framework, LViTES, for improving segmentation of endoscopic images by leveraging both visual and textual information. Our approach builds upon the LViT model but introduces significant enhancements, such as utilizing the EfficientNet backbone for image feature extraction and incorporating a cross-attention mechanism during the decoding process. Additionally, we addressed the lack of textual data by proposing a novel method for generating textual descriptions directly from lesion masks, which enhances the model's ability to generalize even when physician-provided reports are unavailable. We evaluated LViTES on three datasets, including gastric cancer, esophageal cancer, and polyps (Kvasir-SEG), and demonstrated that our model outperforms state-of-the-art methods in terms of segmentation accuracy. Specifically, LViTES showed superior performance in both mIoU and Dice Score metrics compared to the original LViT by around 4% on gastric cancer and esophageal cancer. On Kvasir-SEG dataset, the mIoU of LViTES is higher than TransNetR and the original VLiT by 6%, and ourperformed the U-NET by 11.70%. Overall, LViTES provides a significant step forward in integrating multimodal data for medical image analysis. Future work may explore expanding the scope of this framework to other medical imaging tasks and further improving the quality of generated textual descriptions to boost segmentation performance.

Acknowledgments. This research is funded by Vietnam Ministry of Science and Technology under grant number KC-4.0-17/19-25.

References

1. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE TPAMI **40**(4), 834–848 (2017)
2. Deng, J., Yang, Z., Liu, D., Chen, T., Zhou, W., Zhang, Y., Li, H., Ouyang, W.: Transvg++: End-to-end visual grounding with language conditioned vision transformer. IEEE transactions on pattern analysis and machine intelligence (2023)
3. Devlin, J.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Ding, H., Liu, C., Wang, S., Jiang, X.: Vision-language transformer and query generation for referring segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16321–16330 (2021)
5. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., De Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, proceedings, part II 26. pp. 451–462. Springer (2020)
6. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
7. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A.: H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. IEEE transactions on medical imaging **37**(12), 2663–2674 (2018)
8. Li, Y., Wang, S., Wang, J., Zeng, G., Liu, W., Zhang, Q., Jin, Q., Wang, Y.: Gt u-net: A u-net like group transformer network for tooth root segmentation. In: Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12. pp. 386–395. Springer (2021)
9. Li, Z., Li, Y., Li, Q., Wang, P., Guo, D., Lu, L., Jin, D., Zhang, Y., Hong, Q.: Lvit: language meets vision transformer in medical image segmentation. IEEE transactions on medical imaging (2023)
10. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. pp. 3431–3440 (2015)
11. Luo, X., Chen, J., Song, T., Wang, G.: Semi-supervised medical image segmentation through dual-task consistency. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 8801–8809 (2021)
12. Luo, X., Hu, M., Song, T., Wang, G., Zhang, S.: Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In: International conference on medical imaging with deep learning. pp. 820–833. PMLR (2022)
13. Malaviya, N., Rahevar, M., Virani, A., Ganatra, A., Bhuva, K.: Lvit: Vision transformer for lung cancer detection. In: 2023 International Conference on Artificial Intelligence and Smart Communication (AISC). pp. 93–98. IEEE (2023)
14. Nguyen, T.H., Nguyen, T.D.H., Phung, D., Nguyen, D.T.C., Tran, H.M., Luong, M., Vo, T.D., Bui, H.H., Phung, D., Nguyen, D.Q.: A vietnamese-english neural machine translation system. In: Annual Conference of the International Speech Communication Association (was Eurospeech) 2022. pp. 5543–5544. International Speech Communication Association (ISCA) (2022)
15. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)

16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
18. Shen, S., Li, L.H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.W., Yao, Z., Keutzer, K.: How much can clip benefit vision-and-language tasks? arXiv preprint arXiv:2107.06383 (2021)
19. Tang, S., Yu, X., Cheang, C.F., Liang, Y., Zhao, P., Yu, H.H., Choi, I.C.: Transformer-based multi-task learning for classification and segmentation of gastrointestinal tract endoscopic images. Computers in Biology and Medicine **157**, 106723 (2023)
20. Tran, T.H., Vu, D.H., Tran, D.H., Do, K.L., Nguyen, P.T., Nguyen, V.T., Nguyen, L.T., Ho, N.K., Vu, H., Dao, V.H.: Dcs-unet: Dual-path framework for segmentation of reflux esophagitis lesions from endoscopic images with u-net-based segmentation and color/texture analysis. Vietnam Journal of Computer Science **10**(02), 217–242 (2023)
21. Wang, S., Cong, Y., Zhu, H., Chen, X., Qu, L., Fan, H., Zhang, Q., Liu, M.: Multi-scale context-guided deep network for automated lesion segmentation with endoscopy images of gastrointestinal tract. IEEE Journal of Biomedical and Health Informatics **25**(2), 514–525 (2020)
22. Wang, Y., Huang, Y., Chase, R.C., Li, T., Ramai, D., Li, S., Huang, X., Antwi, S.O., Keaveny, A.P., Pang, M.: Global burden of digestive diseases: a systematic analysis of the global burden of diseases study, 1990 to 2019. Gastroenterology **165**(3), 773–783 (2023)
23. Wu, Y., Ge, Z., Zhang, D., Xu, M., Zhang, L., Xia, Y., Cai, J.: Mutual consistency learning for semi-supervised medical image segmentation. Medical Image Analysis **81**, 102530 (2022)
24. Xu, F., Lin, L., Li, D., Hong, Q., Liu, K., Wu, Q., Li, Q., Zheng, Y., Tian, J.: A multi-resolution deep forest framework with hybrid feature fusion for ct whole heart segmentation. In: International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1119–1124. IEEE (2021)
25. Xu, F., Lin, L., Li, Z., Hong, Q., Liu, K., Wu, Q., Li, Q., Zheng, Y., Tian, J.: Mrdff: A deep forest based framework for ct whole heart segmentation. Methods **208**, 48–58 (2022)
26. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Lavt: Language-aware vision transformer for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18155–18165 (2022)
27. Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
28. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. pp. 3–11. Springer (2018)