

RELIABILITY-GUIDED VISION-LANGUAGE DECODING WITH TEXT-AWARE SKIP REFINEMENT FOR MEDICAL IMAGE SEGMENTATION

Sivan Talaie¹, Fatemeh Daneshfar^{2,*}, Mustaqeem Khan³, Abdulhady Abas Abdullah⁴, Mourad Oussalah⁵

^{1,2} Department of Computer Engineering, University of Kurdistan, Sanandaj, Kurdistan, Iran

³ College of Information Technology, United Arab Emirates University, Al Ain, UAE

⁴ Artificial Intelligence and Innovation Centre, University of Kurdistan, Erbil, Kurdistan, Iraq

⁵ Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, Finland

ABSTRACT

Medical image segmentation remains challenging when targets exhibit limited visual distinction, irregular structure, or substantial variation across imaging modalities. Medical text can reduce this ambiguity, but indiscriminate fusion may introduce unreliable guidance. We present RELiF, a vision-language framework that integrates text through reliability-guided decoding. Its Reliability-guided Fusion Gate (RFG) adaptively controls cross-modal updates, while the Text-guided Skip Refinement Adapter (TSRA) refines high-resolution features using text guidance. RELiF is evaluated on radiographic, CT, and endoscopic segmentation tasks, where it shows competitive performance across diverse imaging characteristics and target appearances. Additional analyses examine efficiency and stability. The complementary uncertainty representation derived from the reliability response also shows an association with prediction difficulty, with higher uncertainty values occurring more frequently in pixels associated with segmentation errors. Code: <https://github.com/SeivanTalaie/RELiF>.

Index Terms— Medical image segmentation, Vision-language learning, Multimodal fusion

1. INTRODUCTION

Medical image segmentation supports clinical assessment, treatment planning, and disease monitoring by identifying anatomical structures and pathological regions. Unimodal approaches, such as U-Net [1], rely solely on visual information, which can limit their effectiveness when targets exhibit subtle appearances or substantial variation across imaging conditions. To address these challenges, multimodal approaches incorporate complementary semantic information from medical text, building on advances in vision-language learning such as CLIP [2]. However, the benefits of textual guidance depend not only on the information it provides but also on how effectively it is integrated with visual features.

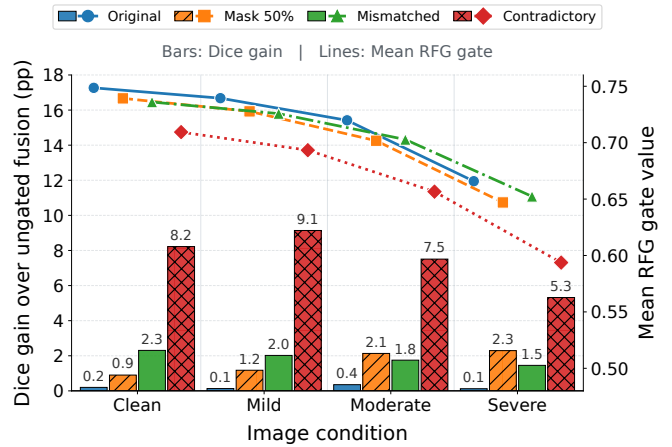


Fig. 1. RFG behavior under controlled image and text degradation on QaTa-COV19. Lines show the mean gate response across the test set and three RFG stages, while bars indicate Dice gains (pp) over ungated fusion.

Recent studies have explored cross-modal reconstruction [3], uncertainty-aware learning [4], and bidirectional fusion [5] to improve vision-language segmentation. However, how to adaptively regulate textual guidance across different visual regions during decoding remains an open question, as its relevance may vary spatially. Another limitation lies in conventional skip connections, which directly transfer high-resolution features without considering their relevance, potentially introducing domain-specific variations into the decoder [6]. These challenges call for selective cross-modal fusion and context-aware refinement of skip features.

To address these challenges, we propose RELiF, a vision-language framework that combines reliability-guided decoding with text-guided skip refinement. Its Reliability-guided Fusion Gate (RFG) estimates token-wise reliability from visual and text-conditioned features, together with their feature-level agreement and discrepancy, to adaptively regulate cross-modal contributions during decoding. As shown in Fig. 1, RFG adapts its fusion response to variations in image qual-

*Corresponding author: f.daneshfar@uok.ac.ir

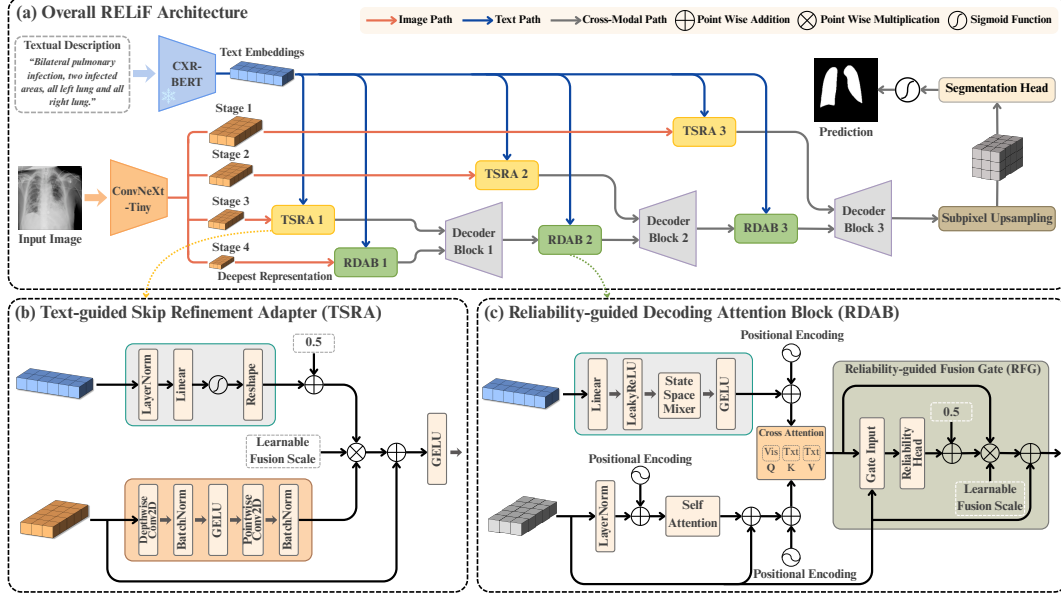


Fig. 2. Overview of RELiF: (a) vision–language segmentation architecture. (b) Text-guided Skip Refinement Adapter (TSRA). (c) Reliability-guided Decoding Attention Block (RDAB) with its Reliability-guided Fusion Gate (RFG).

ity and textual relevance, with corresponding Dice gains over ungated fusion. In parallel, the Text-guided Skip Refinement Adapter (TSRA) uses textual context to refine high-resolution skip features before decoder fusion. The main contributions are as follows:

- We introduce RELiF, whose Reliability-guided Fusion Gate (RFG) within the Reliability-guided Decoding Attention Block (RDAB) adaptively regulates cross-modal fusion using token-wise reliability derived from visual and text-conditioned features and their feature-level similarities and differences.
- We develop the Text-guided Skip Refinement Adapter (TSRA) to preserve fine structural details through text-conditioned channel modulation and lightweight refinement of high-resolution skip features.
- We evaluate RELiF on chest X-ray and lung CT infection segmentation, as well as endoscopic polyp segmentation, demonstrating its effectiveness across diverse imaging modalities and target appearances.
- We conduct additional experiments and analyses to assess RELiF’s effectiveness, reliability, efficiency, and stability.

2. METHODOLOGY

2.1. Overview

As shown in Fig. 2, visual and textual features guide progressive decoding. RDABs regulate cross-modal fusion, while TSRA operates on separate skip pathways, refining encoder features before their integration into decoder blocks. Finally, subpixel upsampling restores the spatial resolution, and the

segmentation head generates the final pixel-wise predictions.

2.2. Reliability-guided Fusion Gate (RFG)

Given medical images $\mathbf{I} \in \mathbb{R}^{B \times C_I \times H \times W}$ and tokenized descriptions $\mathbf{T}$, ConvNeXt-Tiny [7] extracts visual features $\mathbf{I}'_s$ across four encoder stages, $s \in \{1, 2, 3, 4\}$, while BioViL CXR-BERT [8] generates text embeddings $\mathbf{T}' \in \mathbb{R}^{B \times L \times D_T}$. Here, B denotes batch size, C_I the image channels, H, W the spatial dimensions, and L, D_T the text length and embedding dimension. Decoding starts from the deepest visual features $\mathbf{I}'_4$, reshaped into $\mathbf{V} \in \mathbb{R}^{B \times N_4 \times C_4}$, where $N_4 = H_4 W_4$ is the number of visual tokens. Meanwhile, text embeddings pass through a learnable linear layer, LeakyReLU, SSMixer, and GELU to produce $\mathbf{T}_p \in \mathbb{R}^{B \times L_p \times C_4}$, where L_p is the resulting sequence length. SSMixer models dependencies among textual tokens to provide context-aware representations at the current visual scale. Within RDAB, normalized and position-encoded visual tokens undergo multi-head self-attention. The output is normalized and added to $\mathbf{V}$ through a residual connection, yielding contextual features $\mathbf{V}_c$. Cross-attention then uses normalized, position-encoded $\mathbf{V}_c$ as queries, position-encoded $\mathbf{T}_p$ as keys, and $\mathbf{T}_p$ as values to generate text-conditioned features $\mathbf{C}$. To assess their consistency with visual information, RFG combines both representations and their feature-level relationships:

$$\mathbf{A} = \text{Concat}(\mathbf{V}_c, \mathbf{C}, |\mathbf{V}_c - \mathbf{C}|, \mathbf{V}_c \odot \mathbf{C}). \quad (1)$$

A lightweight reliability head then estimates a response for each visual token:

$$\mathbf{R} = \sigma(\text{Linear}(\text{GELU}(\text{Linear}(\text{LN}(\mathbf{A}))))). \quad (2)$$

The complementary uncertainty representation is defined as $\mathbf{U} = 1 - \mathbf{R}$, with higher values indicating lower estimated reliability. The reliability response subsequently regulates cross-modal fusion through

$$\mathbf{G} = 0.5 + \mathbf{R}, \quad \mathbf{Y} = \mathbf{V} + \alpha(\mathbf{G} \odot \mathbf{C}), \quad (3)$$

where $\odot$ denotes element-wise multiplication, σ the sigmoid function, and α a learnable scalar. The bounded gate $\mathbf{G} \in [0.5, 1.5]$ regulates cross-modal contributions without entirely suppressing or excessively amplifying them.

2.3. Text-guided Skip Refinement Adapter (TSRA)

While RFG regulates cross-modal fusion, TSRA refines high-resolution skip features $\mathbf{I}'_s$ for $s \in \{1, 2, 3\}$ to preserve relevant spatial details. The shared text embedding $\mathbf{T}'$ is averaged across L tokens to obtain $\mathbf{t}_g$, which is normalized and projected into a stage-specific channel response. A sigmoid activation and a 0.5 shift are then applied, followed by reshaping to enable channel-wise modulation of the spatial skip features, yielding $\bar{\mathbf{g}}_s \in [0.5, 1.5]^{B \times C_s \times 1 \times 1}$:

$$\bar{\mathbf{g}}_s = \text{Reshape}(0.5 + \sigma(\text{Linear}(\text{LN}(\mathbf{t}_g)))) \quad (4)$$

In parallel, a depthwise 3×3 convolution extracts local spatial details from $\mathbf{I}'_s$, followed by batch normalization and GELU. A pointwise 1×1 convolution then combines information across channels, followed by another normalization to produce $\mathbf{D}_s$. The text-conditioned gate $\bar{\mathbf{g}}_s$ modulates these features through a scaled residual connection:

$$\tilde{\mathbf{I}}'_s = \text{GELU}(\mathbf{I}'_s + \beta(\bar{\mathbf{g}}_s \odot \mathbf{D}_s)) \in \mathbb{R}^{B \times C_s \times H_s \times W_s}. \quad (5)$$

Here, β is a learnable residual scale, and $\odot$ denotes element-wise multiplication. Zero-initialized gate projections and small initial residual scales allow gradual refinement. Finally, the refined skip features $\tilde{\mathbf{I}}'_s$ are fused with the upsampled decoder features without additional supervision.

2.4. Decoder Block

The decoder progressively restores spatial details through three stages, $s \in \{4, 3, 2\}$. At each stage, RDAB uses the shared text embedding $\mathbf{T}'$ to produce reliability-guided visual tokens $\mathbf{Y}_s \in \mathbb{R}^{B \times N_s \times C_s}$, where $N_s = H_s W_s$ denotes the token count and C_s the channel dimension. These tokens are reshaped into spatial features $\mathbf{Z}_s \in \mathbb{R}^{B \times C_s \times H_s \times W_s}$ and upsampled through transposed convolution. The resulting features are concatenated with the corresponding TSRA-refined skip features $\tilde{\mathbf{I}}'_{s-1}$ and convolved:

$$\mathbf{D}_{s-1} = \text{ConvBlock}\left(\text{Concat}\left(\text{UpConv}(\mathbf{Z}_s), \tilde{\mathbf{I}}'_{s-1}\right)\right) \quad (6)$$

The fused features are converted back into visual tokens for the next RDAB, progressively incorporating finer spatial

information. Finally, subpixel upsampling restores the input resolution, and a sigmoid-activated segmentation head generates pixel-wise foreground probabilities:

$$\widehat{\mathbf{M}} = \sigma(\text{SegHead}(\text{SubPixelUp}(\mathbf{D}_1))). \quad (7)$$

2.5. Objective Function

RELIF is trained using the Spectral-Entropic Uncertainty (SEU) objective [4], which combines three complementary loss terms. Dice loss maximizes overlap between the predicted mask $\widehat{\mathbf{M}}$ and ground truth $\mathbf{M}$, while spectral loss encourages structural consistency by comparing their Fourier magnitude spectra. An entropy-inspired term further promotes confident foreground predictions. The overall objective is defined as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Dice}} + \lambda_F \mathcal{L}_{\text{Spectral}} + \lambda_E \mathcal{L}_{\text{Entropy}}, \quad (8)$$

where λ_F and λ_E control the contributions of the spectral and entropy terms, respectively. The entropy-inspired term is applied only to the final segmentation probabilities and is separate from the RFG-derived uncertainty representation. RFG and TSRA are learned end-to-end without direct supervision.

3. EXPERIMENTS AND RESULTS

3.1. Datasets

Experiments were conducted on three public datasets: QaTa-COV19 [9], MosMedData+ [10], and Kvasir-SEG [11], covering chest X-ray, CT, and endoscopic segmentation. Following the LViT protocol [12], we adopt the same dataset splits and use text annotations from [12] for infection segmentation and [13] for polyp segmentation. The dataset splits are: QaTa-COV19: Training=5716; Validation=1429; Test=2113; MosMedData+: Training=2183; Validation=273; Test=273; Kvasir-SEG: Training=800; Validation=100; Test=100.

3.2. Implementation Details

The framework was implemented in PyTorch, with each dataset trained independently on an NVIDIA RTX 4090 GPU. All input images and masks were resized to 224×224 pixels. Training was conducted for 200 epochs, using batch sizes of 32 for training and 8 for validation and testing. We used the AdamW optimizer with a cosine annealing learning-rate scheduler. Initial learning rates were 3×10^{-4} for QaTa-COV19 and Kvasir-SEG and 5×10^{-4} for MosMedData+. The SEU loss weights were set to $\lambda_F = 0.3$ and $\lambda_E = 0.1$.

3.3. Experimental Results and Analysis

Table 1 compares RELIF with representative monomodal and multimodal methods across three medical segmentation datasets. RELIF achieves the highest Dice and mIoU scores

Table 1. Comparison with state-of-the-art methods. best and second-best results are **bold** and underlined, respectively.

Method	QaTa-COV19		MosMedData+		Kvasir-SEG	
	Dice(%)	mIoU(%)	Dice(%)	mIoU(%)	Dice(%)	mIoU(%)
<i>Monomodal models</i>						
U-Net ('15) [1]	78.91	69.32	64.30	50.50	82.33	74.26
nnUNet ('21) [14]	80.30	70.62	72.32	60.14	84.22	75.41
UCTransNet ('22) [15]	79.00	69.34	65.71	52.55	91.04	<u>87.31</u>
<i>Multimodal models</i>						
LViT ('23) [12]	83.40	74.89	74.32	61.03	87.59	75.16
RecLMIS ('25) [3]	84.93	76.86	77.26	64.95	86.58	77.08
ProLearn ('25) [16]	<u>90.60</u>	<u>82.81</u>	77.82	63.69	90.40	82.80
UA-VLS ('26) [4]	90.11	82.00	78.48	64.58	<u>92.41</u>	85.89
BiCLIP ('26) [5]	<u>90.60</u>	<u>82.81</u>	81.00	68.06	89.02	80.21
RELIF (Ours)	91.54	84.41	<u>79.27</u>	<u>65.66</u>	93.28	87.41

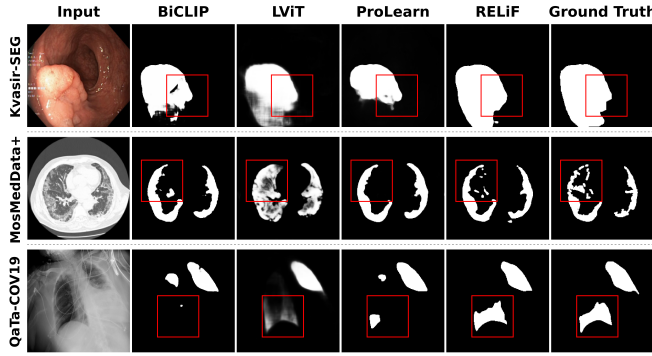


Fig. 3. Qualitative segmentation results across three datasets.

Table 2. Segmentation performance with 50% and 10% of the training data across three datasets.

Method	Regime	QaTa-COV19		MosMedData+		Kvasir-SEG	
		Dice(%)	mIoU(%)	Dice(%)	mIoU(%)	Dice(%)	mIoU(%)
BiCLIP [5]	50%	89.83	81.53	79.72	66.27	86.55	76.29
	10%	87.16	77.25	69.23	52.95	71.72	55.91
LViT [12]	50%	78.81	69.55	58.98	47.28	80.66	73.30
	10%	74.72	64.99	51.07	39.59	55.68	44.38
RELIF	50%	91.10	83.66	78.06	64.01	91.95	85.10
	10%	88.29	79.04	74.72	59.64	85.17	74.17

on QaTa-COV19 and Kvasir-SEG, while maintaining competitive performance on MosMedData+ with only 18.1 G FLOPs. Qualitative comparisons in Fig. 3 further illustrate the segmentation results across all three datasets. To assess robustness under limited supervision, Table 2 compares RELiF with BiCLIP and LViT using 50% and 10% training data. RELiF achieves the best Dice and mIoU on two datasets at 50% and all three at 10%. We further evaluated stability across five independent training runs, obtaining Dice standard deviations of 0.12, 0.14, and 0.24 percentage points on QaTa-COV19, MosMedData+, and Kvasir-SEG, respectively, indicating low run-to-run variability. The ablation study in Table 4 confirms the complementary contributions of RFG and TSRA, with their combination achieving the best performance. Furthermore, removing textual guidance

Table 3. Uncertainty-based pixel error detection using RFG-derived uncertainty response. Error prevalence defines the random AUPRC baseline, while AUPRC lift measures gains.

Dataset	Error prevalence(%)	AUROC $\uparrow$	AUPRC $\uparrow$	Random AUPRC	AUPRC lift $\uparrow$
QaTa-COV19	2.03	0.7724	0.0824	0.0203	4.06 $\times$
MosMedData+	0.65	0.8389	0.0199	0.0065	3.04 $\times$
Kvasir-SEG	1.88	0.8473	0.1035	0.0188	5.51 $\times$

Table 4. Ablation study of RELiF evaluates individual components and textual guidance strategies using Dice and mIoU.

Method	MosMedData+		Kvasir-SEG	
	Dice(%)	mIoU(%)	Dice(%)	mIoU(%)
<i>Component Ablation</i>				
Base	78.48	64.58	92.37	85.82
Base + RFG	78.92	65.18	92.89	86.72
Base + TSRA	79.06	65.37	92.91	86.76
Full RELiF (RFG + TSRA)	79.27	65.66	93.28	87.41
<i>Textual Guidance</i>				
Training w/o Text Guidance	77.78	63.64	88.90	80.02
Training w/ Mismatched Text	78.73	64.92	91.01	83.50

or introducing mismatched text during training reduces segmentation accuracy, highlighting the importance of relevant textual information. Table 3 further evaluates the ability of RFG-derived uncertainty to distinguish erroneous pixels from correct predictions, reporting AUROC values of 0.7724–0.8473 and AUPRC scores 3.04–5.51 times their prevalence-based random baselines. These results suggest consistent error-discrimination ability across the three datasets. Finally, we conducted a pixel-level risk–coverage analysis on Kvasir-SEG by progressively excluding high-uncertainty pixels. At 90% coverage, retained-pixel error risk decreased by 45.2%, with selective Dice reaching 93.87%. Further reducing coverage to 70% increased risk reduction to 71.7%, although selective Dice declined, potentially due to the exclusion of correctly predicted foreground pixels.

4. CONCLUSION

This paper introduced RELiF, combining reliability-guided fusion (RFG) and text-conditioned skip refinement (TSRA) for medical segmentation. The results suggest benefits from complementary mechanisms under limited supervision. Sensitivity to mismatched text indicates that gains depend on relevant guidance rather than text alone. RFG-derived uncertainty helps identify unreliable regions, although excessive rejection may remove correctly segmented foreground. These findings motivate further reliability-aware segmentation research.

5. COMPLIANCE WITH ETHICAL STANDARDS

The study retrospectively used open-access human data, with no ethical approval required per the dataset license.

6. ACKNOWLEDGMENTS

This study received no funding, and the authors have no relevant conflicts of interest to disclose.

7. REFERENCES

- [1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. 2015, vol. 9351 of *Lecture Notes in Computer Science*, pp. 234–241, Springer.
- [2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, “Learning transferable visual models from natural language supervision,” in *Proceedings of the 38th International Conference on Machine Learning*. 2021, vol. 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, PMLR.
- [3] Xiaoshuang Huang, Hongxiang Li, Meng Cao, Long Chen, Chenyu You, and Dong An, “Cross-modal conditioned reconstruction for language-guided medical image segmentation,” *IEEE Transactions on Medical Imaging*, vol. 44, no. 4, pp. 1821–1835, Apr. 2025.
- [4] Aryan Das, Tanishq Rachamalla, Koushik Biswas, Swalpa Kumar Roy, and Vinay Kumar Verma, “Uncertainty-aware vision-language segmentation for medical imaging,” in *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026, pp. 8490–8499.
- [5] Saivan Talaei, Fatemeh Daneshfar, Abdulhady Abas Abdullah, and Mustaqeem Khan, “BiCLIP: Bidirectional and consistent language-image processing for robust medical image segmentation,” *arXiv preprint arXiv:2603.00156*, 2026.
- [6] Frauke Wilm, Jonas Ammeling, Mathias Öttl, Rutger H. J. Fick, Marc Aubreville, and Katharina Breininger, “Rethinking u-net skip connections for biomedical image segmentation,” *arXiv preprint arXiv:2402.08276*, 2024.
- [7] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie, “A convnet for the 2020s,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11976–11986.
- [8] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay, “Making the most of text semantics to improve biomedical vision–language processing,” in *Computer Vision – ECCV 2022*, Cham, Oct. 2022, vol. 13696 of *Lecture Notes in Computer Science*, pp. 1–21, Springer Nature Switzerland.
- [9] Aysen Degerli, Serkan Kiranyaz, Muhammad E. H. Chowdhury, and Moncef Gabbouj, “Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images,” in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 2306–2310.
- [10] Sergey P Morozov, Anna E Andreychenko, Nikolay A Pavlov, AV Vladzimirskyy, Natalya V Ledikhova, Victor A Gomboleviskiy, Ivan A Blokhin, Pavel B Gelezhe, AV Gonchar, and V Yu Chernina, “Mosmeddata: Chest ct scans with covid-19 related findings dataset,” *arXiv preprint arXiv:2005.06465*, 2020.
- [11] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen, “Kvasir-seg: A segmented polyp dataset,” in *International conference on multimedia modeling*. Springer, 2019, pp. 451–462.
- [12] Zihan Li, Yunxiang Li, Qingde Li, Puyang Wang, Dazhou Guo, Le Lu, Dakai Jin, You Zhang, and Qingqi Hong, “Lvit: Language meets vision transformer in medical image segmentation,” *IEEE Transactions on Medical Imaging*, vol. 43, no. 1, pp. 96–107, 2024.
- [13] Kanchan Poudel, Manish Dhakal, Prasiddha Bhandari, Rabin Adhikari, Safal Thapaliya, and Bishesh Khanal, “Exploring transfer learning in medical image segmentation using vision-language models,” *arXiv preprint arXiv:2308.07706*, 2023.
- [14] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein, “nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation,” *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.
- [15] Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R. Zaiane, “UCTransNet: Rethinking the skip connections in U-Net from a channel-wise perspective with transformer,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, pp. 2441–2449, June 2022.
- [16] Shuchang Ye, Usman Naseem, Mingyuan Meng, and Jinman Kim, “Alleviating textual reliance in medical language-guided segmentation via prototype-driven semantic approximation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 22316–22326.