Region-Based Text-Consistent Augmentation for Multimodal Medical Segmentation

Kunyan Cai[1], Chenggang Yan[2], Min He[3], Liangqiong Qu[4], Shuai Wang[2](✉), and Tao Tan[1](✉)

[1] Faculty of Applied Sciences, Macao Polytechnic University, Macao
{p2317017,taotan}@mpu.edu.mo
[2] Hangzhou Dianzi University, Zhejiang, China
{cgyan,shuaiwang}@hdu.edu.cn
[3] Hangzhou Institute of Medicine, Chinese Academy of Sciences, Zhejiang, China
hemin@him.cas.cn
[4] Faculty of Science, The University of Hong Kong, Hong Kong
liangqqu@hku.hk

Abstract. Medical image segmentation is crucial for various clinical applications, and deep learning has significantly advanced this field. To further enhance performance, recent research explores multimodal data integration, combining medical images and textual reports. However, a critical challenge lies in image data augmentation for multimodal medical data, specifically in maintaining text-image consistency. Traditional augmentation techniques, designed for unimodal images, can introduce mismatches between augmented images and text, hindering effective multimodal learning. To address this, we introduce **R**egion-**B**ased **T**ext-**C**onsistent **A**ugmentation (RBTCA), a novel framework for coherent multimodal augmentation. Our approach performs region-based image augmentation by first identifying image regions described in associated text reports and then extracting textual cues grounded in these regions. These cues are integrated into the image, and augmentation is subsequently performed on this modality-aware representation, ensuring inherent text-cue consistency. Notably, the RBTCA's plug-and-play design allows for straightforward integration into existing medical image analysis pipelines, enhancing its practical utility. We demonstrate the efficacy of our framework on the QaTa-Covid19 and our in-house Lung Tumor CT Segmentation (LTCT) datasets, achieving substantial gains, with a Dice coefficient improvement of up to 7.24% when integrated into baseline segmentation models. Our code will be released on https://github.com/KunyanCAI/RBTCA.

Keywords: Medical Image Segmentation · Data Augmentation · Multimodal Learning · Text-Image Consistency.

[1] (✉) Corresponding Author

1 Introduction

Medical image segmentation has become a critical component of modern healthcare, playing an increasingly important role in diagnosis, treatment planning, and disease monitoring across various modalities such as X-ray, CT, and MRI. The application of deep learning techniques has revolutionized this field, leading to remarkable advancements in automated medical image segmentation [2, 3, 10, 13]. To further enhance the performance and clinical applicability of these segmentation systems, recent research has explored the integration of multimodal data [5, 8], particularly combining medical images with associated textual reports like radiology reports and clinical notes [16]. These multimodal data sources offer complementary information and hold significant promise for achieving a more comprehensive and nuanced understanding of medical images, ultimately leading to improved segmentation performance and clinical utility [7, 12, 19, 22].

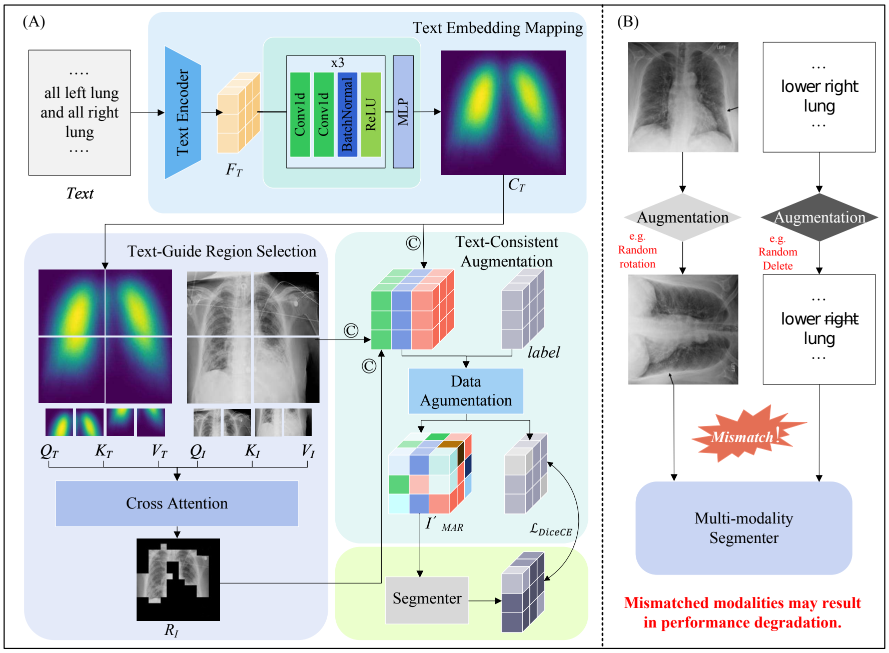

Fig. 1. Overview our proposed RBTCA framework and illustration of multimodal mismatch. (A): Visualized workflow of RBTCA. (B): Illustration of the semantic mismatch problem in multimodal data augmentation.

To effectively train deep learning models, especially in data-scarce domains like medical imaging, data augmentation is an indispensable technique. Data augmentation serves as a crucial strategy to synthetically increase the training dataset by applying various transformations to existing data, thereby mitigating the challenges of limited annotated data, reducing overfitting, and enhancing the generalization capability of deep learning models [4, 14, 18, 23]. This is particularly vital in medical image segmentation, where acquiring large, diverse, and expertly annotated datasets is often hindered by factors such as data scarcity, high annotation costs associated with expert radiologists, and stringent patient privacy regulations [6, 24]. Traditional augmentation strategies, such as geometric transformations (e.g., rotation, flipping) and intensity adjustments (e.g., noise injection, contrast enhancement), have proven effective in unimodal settings by improving model generalization and reducing overfitting [14, 23]. Similarly, in multimodal contexts involving paired image-text data (e.g., radiology reports with X-rays), basic text augmentation techniques like synonym replacement or word deletion are occasionally applied [11, 15]. Yet, these methods operate in isolation, treating image and text modalities as independent streams. As illustrated in Figure 1 (Part B), this isolation introduces semantic inconsistencies: for instance, flipping an X-ray image horizontally while retaining a textual description of "infections in the lower right lung" creates a mismatch between the augmented image and its report. Such discrepancies, quantified in Table 1, degrade the quality of multimodal training data, confuse models during joint representation learning, and ultimately limit segmentation accuracy.

Current augmentation frameworks fail to address a critical question: How can we ensure semantic alignment between augmented images and their associated textual descriptors in multimodal medical data? Most methods either focus solely on images, ignore modality interactions, or require extensive architectural modifications to segmentation networks. This limits their practicality, as medical imaging pipelines often rely on standardized, pre-trained models (e.g., *U-Net* [13]) that cannot be easily redesigned for modality-aware augmentation.

Table 1. Dice (%) for QaTa-Covid19 Segmentation in different augmentation modality.

Model	**Modalities**	**Augmentation Modalities**			
		None	**Image**	**Text**	**Image+Text**
SwinUnet	Image	71.08	74.01	-	-
LViT	Image+Text	76.98	76.70	76.60	76.89

To bridge this gap, we propose Region-Based Text-Consistent Augmentation (RBTCA), a novel, plug-and-play framework designed to harmonize multimodal augmentations while preserving semantic consistency. Unlike prior work, RBTCA operates as a lightweight preprocessing module, requiring no architectural changes to existing segmentation networks. Our approach leverages text-guided regional cues to align augmentations across modalities: (1) it identifies image regions (e.g., anatomical structures) corresponding to textual descriptions (e.g., "right lung"), (2) embeds these cues into the image to create a modality-aware representation (MAR), and (3) applies data augmentations to this modality-aware representation. This ensures that transformations, such as geometric or intensity adjustments (e.g., flipping, noise injection), are consistently applied to both the image and its embedded textual cues, thus eliminating potential semantic mismatches. The key contributions of this work are threefold:

1) **A Plug-and-Play multimodal augmentation framework:** We propose RBTCA, a novel data augmentation strategy specifically designed for multimodal medical data, enabling seamless integration with diverse architectures (e.g., CNNs, Transformers) without structural modifications.

2) **Effortless multimodal consistency with standard augmentation:** RBTCA achieves multimodal consistency by directly leveraging off-the-shelf image augmentation methods, eliminating the need for modality-specific redesign.

3) **Open-source implementation and empirical validation**: RBTCA will be released as an open-source toolkit for adoption and reproducibility. Empirical validation across diverse architectures and datasets (QaTa-COV19, LTCT) demonstrates significant performance gains over baselines, validating its effectiveness and generalizability for multimodal medical image segmentation.

2 Methods

The RBTCA framework, visually detailed in Figure 1 (Part A), achieves text-consistent data augmentation for multimodal medical image segmentation through a streamlined process. Initially, the framework maps a textual description T into a spatially aligned textual prompt $C_T \in \mathbb{R}^{1\times H\times W\times(D)}$. After that, with C_T's integration into the framework, it identifies a text-guided region of interest (ROI) probability map $R_I \in [0,1]^{1\times H\times W\times(D)}$ to precisely localize the image area semantically corresponding to the textual description. Following this, the framework integrates the textual prompt C_T and ROI R_I with the original medical image $I \in \mathbb{R}^{C\times H\times W\times(D)}$, generating a MAR $I_{MAR} \in \mathbb{R}^{C'\times H\times W\times(D)}$ that incorporates text-derived cues. Subsequently, image augmentation techniques $\mathcal{A}$ are applied to I_{MAR} and its label to produce augmented data $(I'_{MAR}, label') = (\mathcal{A}(I_{MAR}), \mathcal{A}(label))$, which is then used to train a segmentation model S_θ, learning a mapping $S_\theta(I'_{MAR}) \to pred'$ from augmented cue-enhanced images to predicted segmentation masks.

2.1 Text-to-Image Space Mapping and Region of Interest Selection

This stage aims to derive two key components from the input textual description T and the medical image I to a textual prompt in the image spatial space, C_T, and a text-guided ROI, R_I.

Textual Prompt Generation ($T \to C_T$) To generate the textual prompt C_T from the input textual description T, we first tokenize T using a pre-trained tokenizer, $Token(\cdot)$, converting the text into a sequence of tokens. Next, we utilize a pre-trained BERT-base-uncased model with its parameters unfrozen, $E_{BERT}(\cdot)$, trained for our task, to encode the tokenized sequence into a contextualized text embedding, $\mathbf{e}_T \in \mathbb{R}^{L \times d_T}$. The text embedding can be computed as:

$$\mathbf{e}_T = E_{BERT}(Token(T)), \tag{1}$$

where d_T is the dimensionality of the BERT embedding, and L is the sequence length of the tokenized text. To map this text embedding into a spatially structured textual prompt, we employ an Embedding Mapping Block, $M_{EMB}(\cdot) : \mathbb{R}^{L \times d_T} \rightarrow \mathbb{R}^{1 \times H \times W \times (D)}$. Specifically, M_{EMB} comprises three sequential layers followed by a Multilayer Perceptron layer at the end. Each layer within M_{EMB} consists of: a 1-dimensional convolutional layer operating across the text embedding dimension (d_T), followed by another 1-dimensional convolutional layer operating across the sequence length dimension (L), Batch Normalization, and ReLU activation. We then reshape the output of M_{EMB} to obtain the textual prompt $C_T \in \mathbb{R}^{1 \times H \times W \times (D)}$:

$$C_T = M_{EMB}(\mathbf{e}_T). \tag{2}$$

Text-Guided Region of Interest Selection ($(\boldsymbol{I}, \boldsymbol{C_T}) \rightarrow \boldsymbol{R_I}$) To extract the text-guided ROI R_I, we employ a Region of Interest Extraction Module, $M_{ROIE}(I, C_T)$, which takes both the medical image I and the textual prompt C_T as inputs. In the implementation of M_{ROIE}, both the medical image I and the textual prompt C_T are first patchified into patch sequences and then projected into Query (Q), Key (K), and Value (V) representations with learnable linear layers. Thus, for the image I and textual prompt C_T, we obtain query, key, and value representations as (Q_I, K_I, V_I) and $(Q_{C_T}, K_{C_T}, V_{C_T})$, respectively.

Subsequently, we utilize a cross-attention mechanism based on Agent attention[9]. Specifically, the average-pooled image query, $A = AvgPool(Q_I)$, acts as the "Agent" matrix. The region of interest R_I is then computed as:

$$R_I = M_{ROIE}(I, C_T) = \sigma(\sigma(AQ_{C_T}^T)\sigma(AK_{C_T}^T)V_{C_T}), \tag{3}$$

where σ denotes the sigmoid activation function. M_{ROIE} allows the textual prompt C_T to attend to spatially relevant locations in the image I, highlighting the region R_I that is semantically most relevant to the text description T. The output $R_I \in [0, 1]^{H \times W \times (D)}$ indicating the region of interest.

2.2 Text-image Consistent Augmentation (TCA)

Following the previous stage, we proceed to create a MAR, I_{MAR}, by integrating the textual prompt C_T and the ROI R_I with the original medical image I. As C_T and R_I are spatially aligned with I, the cue-enhanced image I_{MAR} is readily generated through a simple channel-wise concatenation. Specifically, we concatenate the textual prompt C_T and the region of interest mask R_I as additional channels to the original medical image I.

$$I_{MAR} = I \oplus C_T \oplus R_I, \tag{4}$$

where $\oplus$ denotes channel-wise concatenation.

Consequently, we can directly apply standard image augmentation techniques to the cue-enhanced image I_{MAR} and its corresponding segmentation label, ensuring no conflict arises between the augmented image and its associated textual prompt. Specifically, we employ a set of common image augmentation transformations $\mathcal{A} = \{\text{Random Flip, Random Rotation, CutMix [21]}\}$. For each training sample $(I_{MAR}, label)$, a randomly selected subset $\mathcal{A}_{selected} \subseteq \mathcal{A}$ of these transformations is applied consistently to both I_{MAR} and *label* to generate the augmented data pair $(I'_{MAR}, label')$. The augmentation process is formally represented as:

$$(I'_{MAR}, label') = \mathcal{A}_{selected}(I_{MAR}, label). \tag{5}$$

2.3 Segmentation Model Training and Loss Function

Our framework leverages two key components, the textual prompt C_T and ROI R_I, for modality-aware representation generation. To train these components, we employ Binary Cross-Entropy loss as auxiliary losses ($\mathcal{L}_{aux}$), applying it to both C_T and R_I with respect to the ground truth. This ensures effective learning of text-to-image space mapping and text-guided region identification.

For segmentation model training, we utilize the augmented modality-aware representation I'_{MAR} and their labels. We employ a standard composite loss function for medical image segmentation, $\mathcal{L}_{seg}$, which combines the Dice loss ($\mathcal{L}_{Dice}$) and the Cross-Entropy loss ($\mathcal{L}_{CE}$) with a weighting factor λ. The loss function can be represented as:

$$\mathcal{L}_{seg} = \lambda \mathcal{L}_{Dice} + (1 - \lambda)\mathcal{L}_{CE} \tag{6}$$

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \mathcal{L}_{aux}. \tag{7}$$

3 Experiments and Results

3.1 Experimental Setup

Datasets and Evaluation Metrics. To validate the effectiveness of RBTCA, we utilized two datasets: the QaTa-COV19 dataset and our in-house Lung Tumor CT Segmentation (LTCT) dataset. The QaTa-COV19 dataset, compiled by researchers from Qatar University and Tampere University, comprises 9258 chest X-ray radiographs from COVID-19 patients, featuring expert annotations of COVID-19 lesion regions and text annotations for multimodal research, as released by Li et al. in LViT [12]; these images have a resolution of 224×224 pixels and were used with the dataset splits from LViT. Additionally, our in-house LTCT dataset, consisting of 143 unenhanced CT scans from 95 patients, provides tumor region annotations and diagnostic reports. Originally at a 512×512 xy resolution with variable slice thicknesses, the LTCT dataset's CT images were preprocessed to achieve an isotropic spacing of $1\times1\times1$ mm^3, axially cropped to 320 slices, windowed to $[-700, 900]$ HU, and linearly normalized to [0, 1].

The LTCT dataset was randomly divided into training (120 scans) and testing (23 scans) sets. For quantitative evaluation, we employed the Dice Similarity Coefficient (Dice) and mean Intersection over Union (mIoU) metrics to assess segmentation performance.

Implementation Details. For image segmentation, we utilized six segmenters: unimodal *Unet* [13], *Swin-Unet* [1], *ConvNeXt* [17], and multimodal *LViT* [12], *ASDA* [20], *ReMamber* [19]. All models were trained in a fully supervised manner for 1000 epochs using the AdamW optimizer with a learning rate of 3e-4. To prevent overfitting, we employed early stopping, stopping training after 50 epochs without Dice improvement. The weighting factor λ in the loss function is set to 0.5 based on preliminary validation ($\lambda \in \{0.1,0.5,0.9\}$), and we found that λ=0.5 balanced Dice and CE losses and yielded strong performance. We used batch sizes 64 for QaTa-Covid19. And for the LTCT dataset (3D CT volumes), we used a batch size of 1 during RBTCA modules' training, and 64 for segmenter training on xy-planar slices for segmenter training. To validate the effectiveness of RBTCA in 3D segmentation, we also employed *3D U-Net* [3] as the segmenter on the LTCT dataset with a batch size of 1.

3.2 Results of Segmentation Performance Evaluation

Comparison with Baseline Methods. To evaluate our module's effectiveness, we compared segmentation performance against baseline models. We first established baselines by assessing six segmenters on both datasets: unimodal *Unet*, *Swin-Unet*, *ConvNeXt*, and multimodal *LViT*, *ASDA*, *ReMamber*. Then, we evaluated performance with our module integrated into these baselines to demonstrate improvements. Detailed results are in Table 2.

The results show that integrating RBTCA significantly improved both unimodal and multimodal segmenters on QaTa-Covid19 Dataset. *SwinUnet* showed the largest gain, with Dice increasing by 7.24% to 78.32% and mIoU by 8.39%. *LViT* achieved the best performance, reaching a Dice of 81.19%. On the LTCT dataset, *Unet* with RBTCA showed a substantial Dice increase from 72.16% to 80.54%. Furthermore, using *3D U-Net* as a segmenter also benefited from RBTCA, with Dice improving from 64.67% to 71.08%

Ablation Study. To assess each component's contribution, we performed an ablation study. Table 2 shows segmentation performance with incremental addition of MAR and TCA. For unimodal models, both MAR and TCA improved performance. For example, *Unet* saw a 2.98% Dice increase with MAR and a further 1.53% with TCA. For multimodal models, MAR significantly improved performance (e.g., *LViT* by 4.21% Dice). However, TCA slightly decreased performance for multimodal models due to potential text-image feature mismatch. That could be explain as TCA's consistency fusion not aligned with original text, which by default inputs to the multimodal segmenter, negatively impacting performance. Thus, we recommend MAR + TCA for unimodal integration and primarily MAR for multimodal integration.

Table 2. Results on the QaTa-Covid19 and Lung Tumor CT Dataset. MAR: Modality-Aware Representation; TCA: Text-Consistent Augmentation.

Backbone	**Text**	**Modules**		**QaTa-Covid19**		**Lung Tumor CT**	
		MAR	TCA	mIoU (%) ↑	Dice (%) ↑	mIoU (%) ↑	Dice (%) ↑
U-Net	×	×	×	75.56±0.22	64.89±0.24	72.16±0.23	60.66±0.21
	✓	✓	×	78.54±0.18	67.86±0.22	77.47±0.19	66.23±0.23
	✓	✓	✓	**80.09±0.18**	**69.98±0.21**	**80.54±0.23**	**70.08±0.25**
SwinUnet	×	×	×	71.08±0.23	59.57±0.25	66.25±0.21	52.61±0.21
	✓	✓	×	77.96±0.19	67.15±0.22	71.31±0.22	57.34±0.23
	✓	✓	✓	**78.32±0.20**	**67.96±0.23**	**72.82±0.19**	**59.65±0.22**
ConvNext	×	×	×	73.99±0.21	62.70±0.24	71.80±0.24	58.84±0.23
	✓	✓	×	77.96±0.19	67.15±0.22	72.01±0.21	59.45±0.22
	✓	✓	✓	**78.32±0.20**	**67.96±0.23**	**74.53±0.20**	**62.03±0.23**
LViT	✓	×	×	76.98±0.21	66.43±0.23	74.85±0.17	63.73±0.20
	✓	✓	×	**81.19±0.17**	**71.23±0.20**	**81.08±0.18**	**71.54±0.21**
	✓	✓	✓	80.23±0.18	70.21±0.21	80.11±0.17	70.44±0.23
ASDA	✓	×	×	77.61±0.19	66.65±0.21	74.39±0.18	62.96±0.22
	✓	✓	×	**78.15±0.18**	66.92±0.21	**75.36±0.17**	**63.79±0.21**
	✓	✓	✓	78.09±0.19	**67.21±0.21**	74.98±0.18	62.87±0.22
ReMamber	✓	×	×	79.12±0.18	68.67±0.21	79.57±0.18	68.44±0.21
	✓	✓	×	**81.03±0.17**	**70.91±0.20**	**80.33±0.19**	**71.06±0.22**
	✓	✓	✓	79.24±0.19	69.99±0.22	80.01±0.20	70.81±0.23
3D Unet	×	×	×	-	-	64.67±0.24	52.28±0.26
	✓	✓	×	-	-	65.40±0.23	52.46±0.24
	✓	✓	✓	-	-	**71.08±0.21**	**62.21±0.23**

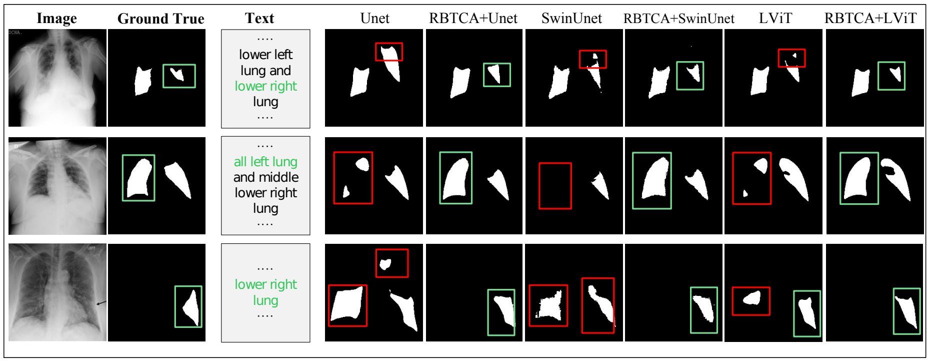
Image
Ground True
Text
Unet
RBTCA+Unet
SwinUnet
RBTCA+SwinUnet
LViT
RBTCA+LViT
.... lower left lung and lower right lung ....
.... all left lung and middle lower right lung ....
.... lower right lung ....

Fig. 2. Visualized result on the QaTa-Covid19 Dataset.

4 Conclusion

This work introduces RBTCA, a novel framework addressing multimodal medical image data augmentation. RBTCA, embeddable by design, achieves coherent augmentation via modality-aware representation and inherent text-image consistency. Key contributions are: (1) a plug-and-play framework for seamless integration, (2) broadly applicable multimodal consistency via standard image augmentation, and (3) significant, generalizable performance gains in segmentation. Extensive evaluations on diverse datasets validate RBTCA's effectiveness in enhancing segmentation accuracy. In summary, RBTCA presents a practical, effective approach to robust multimodal augmentation, contributing to improved deep learning model training and potentially enhanced clinical utility. Future research will explore RBTCA's versatility across medical imaging tasks and modalities.

Acknowledgments. This research was supported by Science and Technology Development Fund of Macao (0041/2023/RIB2), Macao Polytechnic University Grant (RP/FCA-15/2022), National Natural Science Foundation of China under Grant No. 62201323, Zhejiang Provincial Natural Science Foundation of China under Grant No. LDT23F01015F01 and No. LD25F020005, and Basic Research Program of Jiangsu No. BK20220266.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision (pp. 205–218). Cham: Springer Nature Switzerland (2022, October)
2. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
3. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Editor F., Editor S. (eds.) Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19 (pp. 424-432). Springer International Publishing.
4. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation policies from data. arXiv preprint arXiv:1805.09501 (2018)
5. Duan, Y., Pang, P.C.I., He, P., Wang, R., Sun, Y., Liu, C., Zhang, X., Yuan, X., Song, P., Lam, C.T., Cui, L.: 3MT-Net: A Multi-modal Multi-task Model for Breast Cancer and Pathological Subtype Classification Based on a Multicenter Study. IEEE Journal of Biomedical and Health Informatics (2024)
6. Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing **321**, 321–331 (2018)

7. Gao, Y., Zhou, H.Y., Wang, X., Zhang, T., Han, L., Lu, C., Liang, X., Teuwen, J., Beets-Tan, R., Tan, T., Mann, R.: Improving Neoadjuvant Therapy Response Prediction by Integrating Longitudinal Mammogram Generation with Cross-Modal Radiological Reports: A Vision-Language Alignment-Guided Model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 133–143). Cham: Springer Nature Switzerland (2024, October)
8. Guo, Z., Li, X., Huang, H., Guo, N., Li, Q.: Deep learning-based image segmentation on multimodal medical imaging. IEEE Transactions on Radiation and Plasma Medical Sciences **3**(2), 162–169 (2019)
9. Han, D., Ye, T., Han, Y., Xia, Z., Pan, S., Wan, P., Song, S., Huang, G.: Agent attention: On the integration of softmax and linear attention. In: European Conference on Computer Vision (pp. 124–140). Cham: Springer Nature Switzerland (2024, September)
10. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
11. Li, B., Hou, Y., Che, W.: Data augmentation approaches in natural language processing: A survey. Ai Open **3**, 71–90 (2022)
12. Li, Z., Li, Y., Li, Q., et al.: Lvit: language meets vision transformer in medical image segmentation. IEEE transactions on medical imaging **43**(1), 96–107 (2023)
13. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Editor, F., Editor, S. (eds.) Medical image computing and computer-assisted intervention 2015 MICCAI, Munich, Germany, October 5-9, 2015, proceedings, part III 18 (pp. 234–241). Springer International Publishing.
14. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of big data **6**(1), 1–48 (2019)
15. Shorten, C., Khoshgoftaar, T.M., Furht, B.: Text data augmentation for deep learning. Journal of big Data **8**(1), 101 (2021)
16. 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)
17. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16133–16142) (2023)
18. Xiong, X., Sun, Y., Liu, X., Ke, W., Lam, C.T., Chen, J., Jiang, M., Wang, M., Xie, H., Tong, T., Gao, Q.: Distance guided generative adversarial network for explainable medical image classifications. Computerized Medical Imaging and Graphics **118**, 102444 (2024)
19. Yang Y, Ma C, Yao J, et al.: Remamber: Referring image segmentation with mamba twister.In: European Conference on Computer Vision (pp. 108–126). Cham: Springer Nature Switzerland (2024)
20. Yue, P., Lin, J., Zhang, S., Hu, J., Lu, Y., Niu, H., Ding, H., Zhang, Y., Jiang, G., Cao, L., Ji, R.: Adaptive Selection based Referring Image Segmentation. In: Proceedings of the 32nd ACM International Conference on Multimedia (pp. 1101–1110) (2024, October)
21. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision (pp. 6023–6032) (2019)

22. Zhang, Z., Han, L., Zhang, T., Lin, Z., Gao, Q., Tong, T., Sun, Y., Tan, T.: UniMRISegNet: Universal 3D Network for Various Organs and Cancers Segmentation on Multi-Sequence MRI. IEEE Journal of Biomedical and Health Informatics (2024)
23. Zhang H, Cisse M, Dauphin Y N, et al.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
24. Zhao, A., Balakrishnan, G., Durand, F., Guttag, J.V., Dalca, A.V.: Data augmentation using learned transformations for one-shot medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8543–8553) (2019)