Vector-Quantization-Driven Active Learning for Efficient Multi-Modal Medical Segmentation with Cross-Modal Assistance

Xiaofei Du[1,2], Haoran Wang[1,2], Manning Wang[1,2(✉)], and Zhijian Song[1,2(✉)]

[1] Digital Medical Research Center, School of Basic Medical Science, Fudan University, Shanghai 200032, China
[2] Shanghai Key Lab of Medical Image Computing and Computer Assisted Intervention, Shanghai 200032, China
{mnwang, zjsong}@fudan.edu.cn

Abstract. Multi-modal medical image segmentation leverages complementary information across different modalities to enhance diagnostic accuracy, but faces two critical challenges: the requirement for extensive paired annotations and the difficulty in capturing complex inter-modality relationships. While Active Learning (AL) can reduce annotation burden through strategic sample selection, conventional methods suffer from unreliable uncertainty quantification. Meanwhile, Vector Quantization (VQ) offers a mechanism for encoding inter-modality relationships, yet existing implementations struggle with codebook misalignment across modalities. To address these limitations, we propose a novel Vector Quantization - Bimodal Entropy-Guided Active Learning (VQ-BEGAL) framework that employs a dual-encoder architecture with VQ to discretize continuous features into distinct codewords, effectively preserving modality-specific information while mitigating feature co-linearity. Unlike conventional AL methods that separate sample selection from model training, our approach integrates feature-level uncertainty estimation from cross-modal discriminator outputs into the training process—strategically allocating samples with different uncertainty characteristics to optimize specific network components, enhancing both feature extraction stability and decoder robustness.Experiments on benchmark datasets demonstrate that our approach achieves state-of-the-art performance while requiring significantly fewer annotations, making it particularly valuable for real-world clinical applications where labeled data is scarce. The code is available at https://github.com/xf-DU/vq-begal.

Keywords: Multi-modal Medical Image Segmentation · Cross-Modal Assistance · Vector Quantization · Active Learning

1 Introduction

Multi-modal medical image segmentation with cross-modal assistance utilizes auxiliary modalities (e.g., MRI) to assist primary modality (e.g., CT) segmentation, which is critical in computer-aided diagnosis [1]. Although incorporating

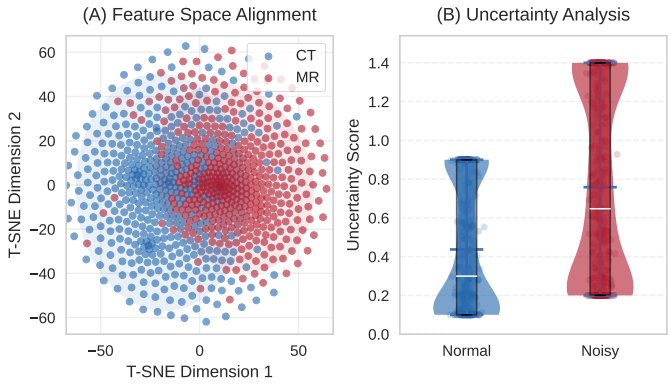

Fig. 1: Key challenges in multi-modal medical image segmentation. **(A)** The t-SNE visualization shows CT and MR features as separated clusters. This is problematic, as shared anatomical features should overlap between modalities while only modality-specific characteristics remain distinct. This improper feature distribution hinders the model's ability to leverage complementary information across modalities. **(B) Uncertainty Analysis:** The uncertainty score distributions between normal and noisy conditions (Gaussian perturbations of input features) reveal how current active learning methods fail in cross-modal settings: their noise sensitivity produces unreliable uncertainty estimates, preventing effective uncertainty estimation for practical multi-modal medical imaging challenges, especially with degraded or noisy modalities.

auxiliary modalities can improve segmentation accuracy, as demonstrated in brain tumor and cardiovascular assessments [1,2], current methods require both modalities during training and inference. This dependency is impractical given the high cost and absence of certain modalities in clinical settings [1,2].

The core challenge lies in effectively disentangling shared anatomical features from modality-specific characteristics while preserving complementary information. Simple multimodal fusion strategies, such as early concatenation, fail to capture nonlinear relationships between modalities, often resulting in the loss of unique complementary information [3,4]. Spatial misalignment and variability in modality quality further exacerbate these issues, as strong linear correlations hinder the model's ability to disentangle shared distinct features [5,6].

Recent work in vector quantization (VQ)-based methods has shown promise in multi-modal feature representation learning by discretizing continuous feature representations into codewords [7,8]. However, as shown in Fig. 1(A), existing VQ approaches often suffer from vector mismatch - where similar anatomical patterns across modalities are encoded with misaligned latent codes - and struggle to disentangle shared anatomical features from modality-specific features, leading to the loss of complementary information. This disentanglement is essential for multi-modal learning, enabling models to leverage common structural patterns while

preserving modality-specific diagnostic details, thus ensuring reliable performance across varying imaging conditions.

These multi-modal representation challenges are exacerbated by limited annotated medical data, making active learning (AL) an effective approach to select informative samples for annotation while maximizing model performance [9]. Although AL reduces annotation costs, conventional strategies [10] face significant limitations in multi-modal settings. As illustrated in Fig. 1(B), these methods yield unreliable uncertainty estimations when modalities are affected by noise, with altered distributions between normal and noisy conditions demonstrating their inability to maintain consistent sample selection in real-world multi-modal scenarios where image quality varies. Furthermore, existing AL approaches typically decouple sample selection from model training, resulting in suboptimal performance as they apply high-uncertainty samples uniformly without addressing the distinct learning objectives of different network components [11,12].

To address these challenges, we propose a novel VQ-BEGAL framework that integrates vector quantization with bimodal entropy-guided active learning for multimodal segmentation. Our dual-encoder architecture uses VQ to discretize continuous feature into distinct codewords, mitigating feature co-linearity while preserving modality-specific details critical for capturing non-linear interactions between modalities. Unlike conventional AL methods that separate sample selection from model training, we incorporate sample selection into the training process itself. We leverage uncertainty estimates from fused multi-modal features to selectively train different network components, utilizing low-uncertainty samples with complementary information to optimize encoder for robustness, while high-uncertainty samples with redundant patterns guide the decoder in capturing modality-specific features. This integrated approach not only reduces labeling costs but also enables more effective multi-modal feature learning.

Our contributions are threefold: (i) we design a dual-encoder architecture with vector quantization that addresses vector mismatch through modality-specific feature extraction and unified feature space learning; (ii) we propose a novel VQ-BEGAL framework that integrates vector quantization for feature disentanglement, paired with an active learning strategy that embeds sample selection directly into the training process and strategically allocates different uncertainty samples to train specific network components; (iii) we conduct extensive experiments on two public datasets, demonstrating that our approach outperforms state-of-the-art methods across various multi-modal segmentation benchmarks.

2 Methodology

2.1 VQ-BEGAL Framework Overview

Our VQ-BEGAL framework features a dual-encoder architecture (Fig. 2) that processes multi-modal inputs through specialized encoders. The framework implements active learning by selecting samples based on discriminator-derived uncertainty scores. Higher uncertainty scores indicate discriminator difficulty in

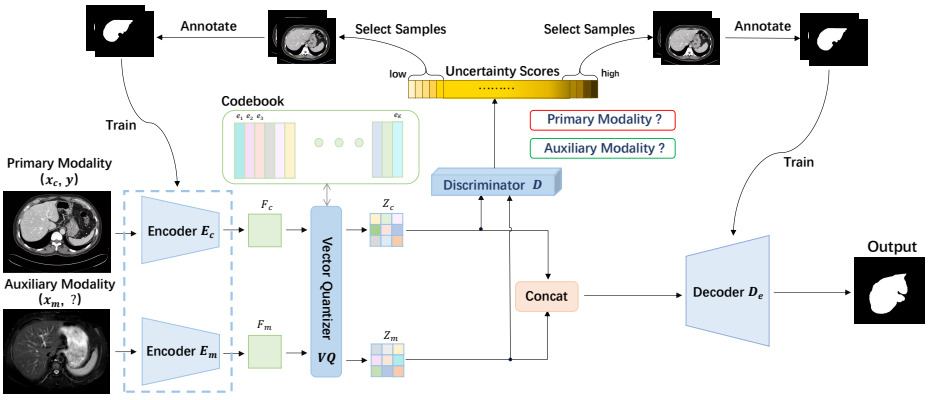

Fig. 2: Overview of our VQ-BEGAL framework. The architecture processes multi-modal inputs through specialized encoders (E_c, E_m) and vector quantization (VQ). The discriminator (D) generates uncertainty scores guiding active learning sample selection (top path), while the main segmentation workflow (bottom path) produces outputs from concatenated features.

distinguishing modalities, suggesting redundant information across them, guiding low-uncertainty samples toward encoder optimization. Conversely, lower uncertainty scores reflect confident predictions with potentially complementary modal information, allocating high-uncertainty samples to enhance decoder robustness. Notably, unlike traditional multi-modal methods, our approach requires no spatial correspondence between modalities, making it more flexible for real-world clinical applications.

2.2 Cross-Modal Auxiliary Feature Learning with VQ

We implement a feature learning strategy with shared VQ between the primary and auxiliary modalities to effectively disentangle shared anatomical features from modality-specific characteristics. The proposed strategy addresses the feature co-linearity and vector mismatch challenges identified in Fig. 1(A).

Dual-Stream Feature Extraction: Modality-specific encoders extract complementary features from primary (e.g., CT) and auxiliary (e.g., MRI) modalities:

$$F_c = E_c(x_c), \quad F_m = E_m(x_m) \tag{1}$$

where $F_c \in \mathbb{R}^{C\times H\times W}$ and $F_m \in \mathbb{R}^{C\times H\times W}$ denote CT and MRI feature maps respectively, with C channels and spatial dimensions $H \times W$. E_c and E_m represent the primary modality and auxiliary modality encoders respectively, while x_c and x_m are the corresponding input images.

Enhanced Vector Quantization: We utilize a codebook with $K = 1024$ entries, replacing standard Euclidean distance with cosine similarity to better capture anatomical feature relationships:

$$d(z, e_k) = \frac{z \cdot e_k}{\|z\| \, \|e_k\|} \tag{2}$$

where z represents an input feature vector, e_k is the k-th codebook entry, and $\frac{z \cdot e_k}{\|z\| \, \|e_k\|}$ measures their similarity.

The quantized features are obtained through:

$$z_c = \mathrm{VQ}(F_c), \quad z_m = \mathrm{VQ}(F_m) \tag{3}$$

where z_c and z_m are the quantized representations of CT and MRI features, VQ represents the vector quantizer.

2.3 Bimodal Entropy-Guided Active Learning

We propose Bimodal Entropy-Guided Active Learning (BEGAL), integrating sample selection into the training process to overcome unreliable uncertainty estimation shown in Fig. 1(B). We leverage a discriminator-based approach to estimate uncertainty and select samples for annotation.

Discriminator Architecture: The discriminator D is implemented as a binary classification network that determines whether the vector-quantized features come from the primary modality or auxiliary modality:

$$p = D(z_c, z_m) \tag{4}$$

where p is the discriminator's predicted probability for each modality class, D represents the discriminator, and z_c and z_m are the quantized features

Uncertainty Estimation: After the discriminator processes the quantized features z_c and z_m from each modality, we compute the uncertainty score based on the output distribution:

$$\mathcal{S}_{uncertainty}(x_c, x_m) = \mathcal{H}(p) = -\sum_{i=1}^{C} p_i \log p_i \tag{5}$$

where $\mathcal{H}(p)$ represents the entropy of the probability distribution, and $\mathcal{S}_{uncertainty}$ is the resulting uncertainty score.

Bimodal Sample Selection: We select both high and low uncertainty samples for different training purposes based on uncertainty as an indicator of cross-modal information redundancy. High uncertainty samples, where the discriminator struggles to distinguish between modalities, contain redundant information that provides stable and consistent training signals ideal for decoder optimization. Low uncertainty samples, with confident discriminator predictions, contain rich complementary cross-modal information suitable for encoder training:

$$\mathcal{S}_{high} = \arg \max_{\mathcal{S} \subset \mathcal{U}, |\mathcal{S}|=n} \sum_{(x_c, x_m) \in \mathcal{S}} \mathcal{S}_{uncertainty}(x_c, x_m) \tag{6}$$

$$\mathcal{S}_{low} = \arg \min_{\mathcal{S} \subset \mathcal{U}, |\mathcal{S}|=n} \sum_{(x_c, x_m) \in \mathcal{S}} \mathcal{S}_{uncertainty}(x_c, x_m) \tag{7}$$

where $\mathcal{U}$ is the unlabeled pool of samples, and $n = \frac{B}{2}$ is determined by dividing the annotation budget B at each active learning round equally between high and low uncertainty sample sets.

High-uncertainty samples ($\mathcal{S}_{high}$) with complementary cross-modal information train the decoder (D_e), while low-uncertainty samples ($\mathcal{S}_{low}$) with redundant information stabilize encoder training (E_c and E_m).

Budget Management: The labeled set $\mathcal{L}$ expands with newly selected samples:

$$\mathcal{L} = \mathcal{L} \cup (\mathcal{S}_{high} \cup \mathcal{S}_{low}) \tag{8}$$

updating the budget:

$$b = b + 2n \tag{9}$$

where b tracks spent annotations against total budget B.

The active learning process terminates when either Dice score plateaus or budget B is exhausted.

3 Experiments

3.1 Datasets

We evaluate our method on two widely-used multi-modal medical image datasets:

CHAOS [13]: The Combined Healthy Abdominal Organ Segmentation dataset comprises 40 paired CT-MRI scans, with expert annotations for the liver, kidneys, and spleen.

AMOS 2022 [14]: The Abdominal Multi-Organ Segmentation dataset consists of 500 CT and 100 MRI scans acquired from multiple medical centers.

We focus on liver segmentation to reduce computational burden and ensure consistent cross-dataset evaluation. Liver segmentation represents a clinically relevant yet challenging task due to variations in contrast, texture, and anatomical boundaries across imaging modalities, making it suitable for validating our multi-modal feature learning approach.

3.2 Implementation Details

We implemented our framework using PyTorch with a VQ-VAE architecture. Our active learning strategy independently selects 50 2D slices extracted from 3D patient data for encoder training and another 50 slices for decoder training in each round, continuing for 10 rounds. Our training objective combines multiple loss components with balanced weights: segmentation loss ($\alpha_1 = 5$), vector

quantization loss ($\alpha_2 = 0.5$), discriminator loss ($\alpha_3 = 0.25$), and commitment loss ($\alpha_4 = 0.2$). The higher weight on segmentation loss ensures task-specific performance while other components provide effective regularization for multi-modal feature learning.

3.3 Comparison with State-of-the-art Methods

We compare our method with several state-of-the-art active learning approaches. As shown in Table 1, our method consistently outperforms existing approaches. The improvement stems from our dual-encoder architecture with vector quantization that addresses vector mismatch while preserving modality-specific information, and our discriminative feature learning strategy.

Table 1: Comparison of different active learning methods for liver segmentation on CHAOS and AMOS datasets. Results are reported with 40% annotation budget.

Method	CHAOS		AMOS	
	Dice (%)	HD95 (mm)	Dice (%)	HD95 (mm)
CT-only	78.25±1.25	13.15±0.95	77.92±1.30	13.45±0.98
Random	79.45±1.23	12.82±0.92	78.92±1.28	13.12±0.95
Max Entropy [15,16]	80.12±1.15	11.45±0.85	79.65±1.20	12.85±0.88
MC Dropout [17]	81.85±1.08	10.21±0.78	80.18±1.15	11.52±0.82
Coreset [18]	82.24±1.12	9.94±0.82	81.75±1.10	10.28±0.80
BADGE [19]	83.68±1.05	9.72±0.75	82.12±1.08	9.95±0.76
TAAL [20]	84.95±1.02	9.58±0.72	83.45±1.05	9.68±0.73
MVAAL [21]	85.02±1.04	8.83±0.67	84.02±0.99	8.79±0.77
BEGAL (Ours)	**87.30±0.95**	**8.21±0.68**	**85.43±0.98**	**8.35±0.70**

HD95: 95th percentile Hausdorff Distance

3.4 Ablation Studies

To validate the effectiveness of our proposed method, we conduct ablation studies from three aspects:

Table 2: Ablation study on different components under various annotation ratios. Results are reported in Dice score (%).

Method	20%	30%	40%	50%
Baseline (U-Net) [22]	75.45	78.82	81.25	83.48
Baseline+EGAL	77.68	80.45	83.82	85.65
Baseline+VQ+Random	78.92	81.25	84.45	86.18
Baseline+BEGAL	79.45	82.85	85.78	87.45
Baseline+VQ (512 codes)+BEGAL	80.82	83.15	86.05	88.68
Baseline+VQ+BEGAL (Ours)	**82.25**	**84.45**	**87.30**	**89.15**

Quantitative Analysis: Table 2 shows each component's contribution to the performance. We use the standard U-Net [22] architecture as our baseline segmentation model. Adding Entropy-Guided Active Learning (EGAL) to the baseline yields a consistent improvement of approximately 2.2-2.6% in Dice score across all annotation ratios, demonstrating the value of entropy-guided active learning. The EGAL differs from our BEGAL approach by using only the highest uncertainty samples for end-to-end model training rather than separately optimizing encoder and decoder components. When we incorporate VQ with random sampling, we observe a further improvement of 1.2-1.5%, highlighting the effectiveness of our discrete representation learning. The combination of baseline with BEGAL alone shows a substantial gain of 3.5-4.5% over the baseline method. Notably, when both VQ and BEGAL are integrated (our full method), we achieve the highest performance with substantial improvements of 5.6-6.8% over the baseline, indicating strong synergy between our discrete representation learning and bidirectional entropy-guided active learning components.

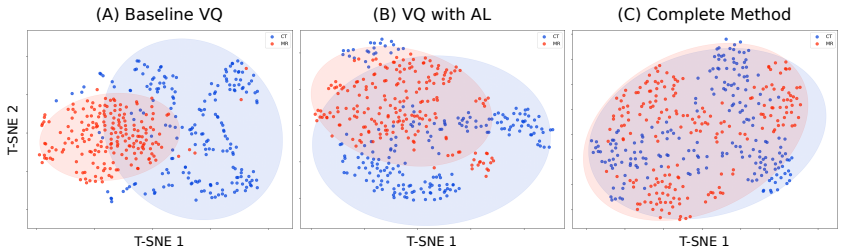

Fig. 3: t-SNE visualization of the quantized representations of CT and MRI features z_c and z_m: (A) Baseline VQ shows limited overlap, (B) VQ with EGAL improves alignment, and (C) our complete method achieves optimal integration.

Quantized Feature Distribution Analysis: Fig. 3 demonstrates how our dual-encoder architecture with VQ addresses vector mismatch. The baseline (A) shows distinct CT and MRI clusters with minimal overlap. Our complete method (C) achieves balanced feature distribution, creating a unified feature space while preserving modality-specific information.

Discriminative Feature Distribution Analysis: Fig. 4 shows our approach effectively separates and utilizes shared and modality-specific patterns. This validates our second contribution of effectively disentangling shared and modality-specific features, producing reliable uncertainty estimates and enable diverse sample selection across modalities.

4 Conclusion

We presented a novel VQ-BEGAL framework that synergistically integrates vector quantization and active learning to address key challenges in multi-modal medical

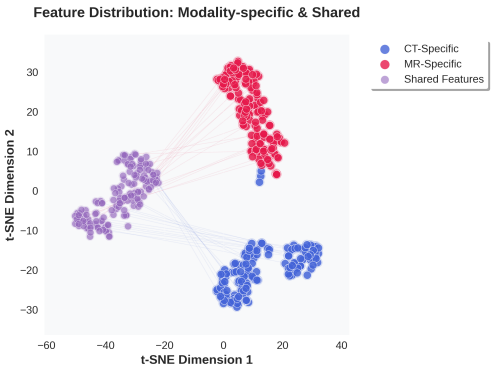

Fig. 4: Discriminative feature distribution across modality conditions, showing BEGAL's ability to effectively disentangling shared and modality-specific features.

image segmentation. Through enhanced feature representation and integrated discriminator-guided sample selection, our method improves training effectiveness while reducing annotation requirements, demonstrating superior performance on two public dataset.

Acknowledgments. No acknowledgments or competing interests.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article[3].

References

1. Zhang, Y., Yang, J., Tian, J., Shi, Z., Zhong, C., Zhang, Y., He, Z.: Modality-aware mutual learning for multi-modal medical image segmentation. In: MICCAI 2021. LNCS, vol. 12901, pp. 589–599. Springer, Heidelberg (2021)
2. Ting, H., Liu, M.: Multimodal transformer of incomplete MRI data for brain tumor segmentation. IEEE Journal of Biomedical and Health Informatics **28**(1), 89–99 (2023)
3. Lu, J., Chen, J., Cai, L., Jiang, S., Zhang, Y.: H2ASeg: Hierarchical Adaptive Interaction and Weighting Network for Tumor Segmentation in PET/CT Images. In: MICCAI 2024. LNCS, pp. 316–327. Springer (2024)
4. Long, L., Cui, J., Zeng, P., Li, Y., Liu, Y., Wang, Y.: MuGI: Multi-Granularity Interactions of Heterogeneous Biomedical Data for Survival Prediction. In: MICCAI 2024. LNCS, pp. 490–500. Springer (2024)
5. Zhou, T., Ruan, S., Canu, S.: A review: Deep learning for medical image segmentation using multi-modality fusion. Array **3**, 100004 (2019)

[3] If EquinOCS, our proceedings submission system, is used, then the disclaimer can be provided directly in the system.

6. Yuan, M., Wei, X.: C 2 Former: Calibrated and Complementary Transformer for RGB-Infrared Object Detection. IEEE Transactions on Geoscience and Remote Sensing (2024)
7. Kolesnikov, A., Pinto, A.S., Beyer, L., Zhai, X., Harmsen, J., Houlsby, N.: Uvim: A unified modeling approach for vision with learned guiding codes. In: Advances in Neural Information Processing Systems **35**, pp. 26295–26308 (2022)
8. Gorade, V., Mittal, S., Jha, D., Bagci, U.: Synergynet: Bridging the gap between discrete and continuous representations for precise medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 7768–7777 (2024)
9. Wang, H., Du, X., Wang, M., Yao, F., Zhang, L., Song, Z.: A comprehensive survey on deep active learning in medical image analysis. Medical Image Analysis **94**, 103139 (2024)
10. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: ICML 2016, pp. 1050–1059. PMLR (2016)
11. Gaillochet, M., Desrosiers, C., Lombaert, H.: Active learning for medical image segmentation with stochastic batches. Medical Image Analysis **90**, 102958 (2023)
12. Shi, J., Ruan, S., Zhu, Z., Zhao, M., An, H., Xue, X., Yan, B.: Predictive Accuracy-Based Active Learning for Medical Image Segmentation. arXiv preprint arXiv:2405.00452 (2024)
13. Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., Baydar, B., Lachinov, D., Han, S., Pauli, J., Isensee, F., Andrearczyk, V., Müller, H., Menze, B.H., Maier-Hein, K.H., Selver, M.A.: CHAOS Challenge - Combined (CT-MR) Healthy Abdominal Organ Segmentation. Medical Image Analysis **69**, 101950 (2021)
14. Ji, Y., Tang, Y., Gao, R., Zhang, L., Gao, Y., Mei, X., Zhang, J., Li, Q., Zhang, Y.D., Yuille, A.L.: AMOS: A Large-Scale Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation. IEEE Transactions on Medical Imaging **42**(5), 1556–1570 (2023)
15. Shannon, C.E.: A Mathematical Theory of Communication. Bell System Technical Journal **27**(3), 379–423 (1948)
16. Settles, B.: Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison (2009)
17. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. International Conference on Machine Learning (ICML), 1050–1059 (2017)
18. Sener, O., Savarese, S.: Active Learning for Convolutional Neural Networks: A Core-Set Approach. International Conference on Learning Representations (ICLR) (2018)
19. Ash, J.T., Zhang, C., Krishnamurthy, A., Langford, J., Agarwal, A.: Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. International Conference on Learning Representations (ICLR) (2020)
20. Zhang, C., Yao, Y., Zhang, H., Tao, D.: Task-Aware Active Learning for Semantic Segmentation. European Conference on Computer Vision (ECCV), 330–346 (2020)
21. Khanal, B., Bhattarai, B., Khanal, B., Stoyanov, D., Linte, C.A.: M-vaal: Multimodal variational adversarial active learning for downstream medical image analysis tasks. Annual Conference on Medical Image Understanding and Analysis, 48–63 (2023)
22. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 234–241 (2015)