Organ Level Representation Learning for Region Based Medical Image Retrieval

Donghwan Lee
Department of Industrial Engineering
Yonsei University
Seoul, Republic of Korea
dhlee.ie@yonsei.ac.kr

Wooju Kim*
Department of Industrial Engineering
Yonsei University
Seoul, Republic of Korea
wkim@yonsei.ac.kr

Abstract

As medical image databases expand, precise Content-Based Medical Image Retrieval (CBMIR) techniques are increasingly required to support case-based reasoning, clinical education, and data-driven decision-making. Recent deep learning–based CBMIR approaches typically rely on global embeddings to enhance retrieval performance. However, such image-level representations often dilute localized anatomical features and fail to capture clinically relevant organ-specific details. To address this limitation, we propose a region-based CBMIR framework that integrates organ-level information into both representation learning and retrieval. The ROI Embedding Selector extracts patch-level embeddings from user-specified regions of interest (ROIs). The Region-aware Organ Attention (ROA) module then learns structured organ representations through cross-attention between image patches and dedicated organ tokens. During inference, a visibility-weighted aggregation strategy guided by Organ Visibility Recognition incorporates query-relevant organs, enabling anatomically targeted and clinically meaningful retrieval. Experiments on the TotalSegmentator dataset demonstrate that the proposed framework consistently outperforms global embedding–based vision foundation models, particularly in region query settings.

1 Introduction

Medical images play an important role in clinical practice, including diagnosis, treatment planning, and prognosis prediction. Advances in imaging modalities such as computed tomography (CT), positron emission tomography (PET), magnetic resonance imaging (MRI), and X-ray imaging have enabled more precise assessment of disease severity and progression. These developments contribute to improved patient outcomes and reduced healthcare costs, and provide essential evidence for clinical decision-making (1; 2). Meanwhile, the digitization of healthcare systems and ongoing improvements in imaging devices have accelerated the large-scale accumulation of medical data (3), increasing the need for more precise and reliable methods to organize, manage, and utilize these resources (4).

Medical image retrieval (MIR) has been studied as a technology to support case-based reasoning and clinical education, and to assist disease prediction by artificial intelligence models (5). In particular, Content-Based Medical Image Retrieval (CBMIR), which retrieves images based on visual similarity, has received attention as an approach for large-scale medical image databases. CBMIR enables clinicians to identify visually similar prior cases, supporting clinical decision-making (6).

Over the past decade, advances in deep learning–based representation learning have significantly improved the performance of medical image retrieval. For example, CNN and Vision Transformer–based

*Corresponding author

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

models have achieved substantial performance gains by learning high-level semantic representations (7; 8; 9; 10). Nevertheless, most existing CBMIR methods still rely on global image-level representations, which do not explicitly capture clinically meaningful anatomical structures. As a result, they remain limited in performing fine-grained, region-specific retrieval.

In medical images, diagnosis depends on the anatomical context of lesions and their associated organs. Therefore, global representations alone are insufficient to identify clinically relevant similar cases. Even when a Region of Interest (ROI) is available, global embedding–based methods have limited ability to emphasize local regions or to perform organ-aware, weighted retrieval. To address these limitations, we propose a region-aware, organ-centric CBMIR framework that incorporates anatomical structure into representation learning and retrieval. First, the ROI Embedding Selector extracts patch embeddings corresponding to user-specified regions to enhance local feature representation. Next, the Region-aware Organ Attention (ROA) module refines interactions between image patches and organ tokens to learn structured organ-level representations. During inference, we introduce an Organ Visibility Recognition–based visibility-weighted aggregation strategy to emphasize query-relevant organs and support region-focused retrieval. Through the use of organ-aware representation learning and a visibility-weighted aggregation strategy, the proposed framework mitigates the limitations of global embeddings and performs region-based CBMIR. In addition, experiments on the TotalSegmentator dataset demonstrate that the proposed framework achieves competitive performance in region-based retrieval compared with global embedding–based methods. The main contributions of this work are summarized as follows:

- We propose a region-aware, organ-centric CBMIR framework that integrates anatomical structure into representation learning and retrieval.
- We design an organ-aware representation learning scheme with a visibility-weighted aggregation strategy to enable precise region-based retrieval.
- We validate the proposed framework on the TotalSegmentator dataset and demonstrate strong performance in region-based retrieval compared with global embedding–based methods.

2 Related Work

2.1 Content-Based Medical Image Retrieval

Content-Based Medical Image Retrieval (CBMIR) extracts feature representations from medical images and retrieves similar cases based on feature similarity, where representation quality directly affects retrieval performance. Early methods relied on hand-crafted features and similarity computation in feature space (11; 12). With the advancement of deep learning, convolutional neural networks (CNNs) (13) enabled hierarchical semantic representation learning and improved retrieval performance. Recent studies adopt CNN-based foundation models such as Inception V3 (14), VGG19 (15), and ResNet (16) for feature extraction (7; 8; 9). Lo et al. (2025) (17) further introduced a multi-level CNN framework to model modality–organ–disease relationships. More recently, Transformer-based models (10) have been applied to capture global contextual relationships. Vision Transformers with contrastive learning (18) and attention-based fusion frameworks such as DaRF (5) have been proposed to enhance representation learning. In addition, pretrained vision foundation models, including BiomedCLIP (19), have demonstrated strong retrieval performance without extensive task-specific fine-tuning. Despite these advances, most CBMIR methods rely primarily on global image representations and do not explicitly model anatomical structures for region-specific retrieval.

2.2 Representation Learning using Query Token

Following Vaswani et al. (2017) (10), an increasing number of studies have employed learnable tokens or queries to encode information from data distributions into structured token representations. DETR (20) introduced learnable object queries for object detection, enabling the decoder to perform set prediction by decoding objects from image features through cross-attention. Perceiver (21) and Perceiver IO (22) utilized learnable latent tokens as a bottleneck to project large-scale multimodal inputs into fixed-size latent representations and extended to diverse output formats through query-based readout. VCT (23) learned disentangled concept tokens from image tokens for scene decomposition and representation disentanglement. BoQ (24) employed learnable global queries to aggregate place-specific attributes into a unified representation, improving visual place recognition and retrieval

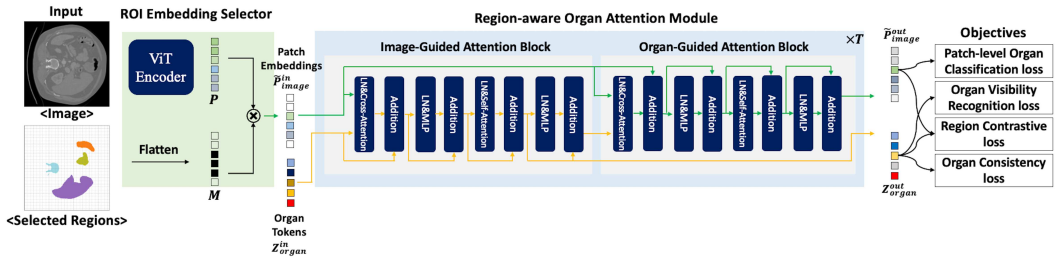

Figure 1: Overall framework of the proposed region-aware organ representation learning framework.

performance. In the medical domain, OWT (25) introduced organ-wise tokenization, decomposing holistic representations into organ-specific token groups and enabling organ-level representation learning through selective token group combinations. However, existing token-based approaches do not explicitly model anatomically meaningful structures or address region-specific retrieval in medical images.

3 Methodology

3.1 ROI Embedding Selector

The ROI Embedding Selector applies a binary mask to retain patch embeddings corresponding to a user-specified region of interest (ROI), ensuring anatomically localized feature extraction (Figure 1). Given patch embeddings $P = p_1, \ldots, p_N$ extracted by a Vision Transformer (ViT) encoder, a binary mask $M \in 0, 1^N$ is defined such that $M_i = 1$ if the i-th patch belongs to the ROI and $M_i = 0$ otherwise. The masked embeddings are computed as $\tilde{p}_i = M_i \cdot p_i$, allowing only ROI patches to contribute to subsequent representation learning. During training, organ-level ROIs are constructed from ground-truth segmentation masks with random region selection to prevent overfitting to specific anatomical structures and promote generalization. During inference, users can manually specify the ROI or select from pre-segmented regions, enabling interactive and region-focused retrieval.

3.2 Region-aware Organ Attention Module

The Region-aware Organ Attention (ROA) module models organ regions in medical images and learns organ-level representations that reflect structural and semantic characteristics. As illustrated in Figure 1, the module takes the masked patch embeddings $\tilde{P}^{\text{in}}_{\text{image}} \in \mathbb{R}^{N \times d}$ obtained from the ROI Embedding Selector and predefined organ token embeddings $Z^{\text{in}}_{\text{organ}} \in \mathbb{R}^{C \times d}$ as inputs. The ROA consists of an Image-Guided Attention (IGA) block and an Organ-Guided Attention (OGA) block, which are alternately applied for T iterations to iteratively refine image and organ representations. Through this interaction, organ tokens aggregate structural cues from masked patch embeddings, while patch representations incorporate organ-level semantic context.

In the IGA block, cross-attention is performed with organ tokens as queries and image patch embeddings as keys and values. A Bin Token is introduced to absorb patch information that does not clearly correspond to a specific organ, reducing interference among organ tokens. Self-attention is then applied among organ tokens to model inter-organ relationships and maintain contextual consistency. In the OGA block, cross-attention is performed with image patch embeddings as queries and organ tokens as keys and values. A Background Token is introduced to separately model background information and stabilize organ-centric representation learning. Self-attention is subsequently applied to image patch embeddings to refine inter-patch relationships and improve structural coherence of the learned representations.

3.3 Objectives

We define a unified training objective consisting of four complementary loss terms for organ-aware representation learning, jointly optimizing spatial accuracy, organ visibility modeling, representation discriminability, and semantic stability.

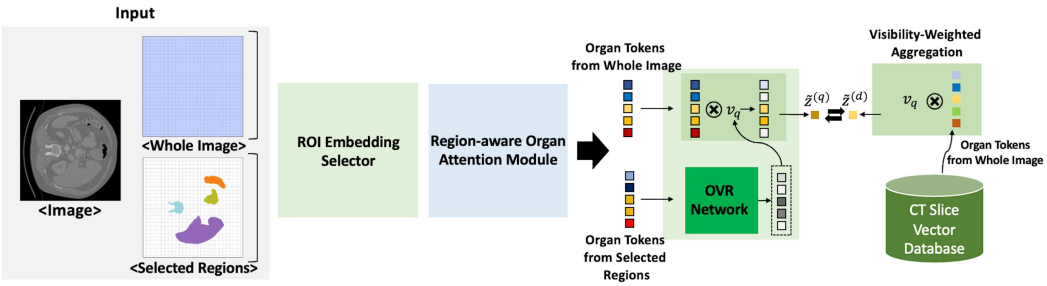

Figure 2: Region-based organ-level retrieval pipeline. Organ tokens are precomputed for database images and matched with query representations based on user-specified regions.

Patch Classification Loss. To provide fine-grained spatial supervision, we introduce a Patch-Level Organ Classification (POC) network that consists of a two-layer MLP. Given the output patch embedding $\tilde{p}_i^{\text{out}}$, the predicted organ label is computed as $\hat{y}_i = \text{POC}(\tilde{p}_i^{\text{out}})$. The supervision is formulated using cross-entropy, defined as $\mathcal{L}_{\text{PC}} = \frac{1}{N}\sum_{i=1}^{N} \text{CE}(y_i, \hat{y}_i)$.

Organ Visibility Loss. To explicitly model organ presence, we introduce an Organ Visibility Recognition (OVR) network, which consists of a two-layer MLP. Given the refined organ token z_c^{out}, the predicted visibility score is computed as $\hat{v}_c = \text{OVR}(z_c^{\text{out}})$. The visibility supervision is formulated using binary cross-entropy, defined as $\mathcal{L}_{\text{OV}} = \frac{1}{C}\sum_{c=1}^{C} \text{BCE}(v_c, \hat{v}_c)$. As shown in Figure 2, the OVR network is used during the image retrieval process.

Region Contrastive Loss. To align organ token embeddings with their corresponding region representations, we introduce a region-level contrastive learning objective. For each organ c, the region embedding r_c is obtained by average pooling the refined patch embeddings $\tilde{p}_i^{\text{out}}$ within the corresponding organ region. The contrastive objective is defined using an InfoNCE-based loss as $\mathcal{L}_{\text{RC}} = -\frac{1}{C}\sum_{c=1}^{C} \log \frac{\exp(\text{sim}(z_c^{\text{out}}, r_c)/\tau)}{\sum_{k=1}^{K} \exp(\text{sim}(z_c^{\text{out}}, r_k)/\tau)}$, where $\{r_k\}_{k=1}^{K}$ denotes region embeddings sampled from all organs within the current mini-batch.

Organ Consistency Loss. To prevent unnecessary updates when relevant evidence is limited, we regularize refined organ embeddings toward their initial organ queries. This keeps unsupported tokens close to their initial states. Given the refined organ embedding z_c^{out} and the corresponding initial query z_c^{in}, the consistency supervision is formulated using an ℓ_2 regression loss. To prevent gradients from flowing into the initial queries, we apply a stop-gradient operator $\text{sg}(\cdot)$, and define the loss as $\mathcal{L}_{\text{OC}} = \frac{1}{C}\sum_{c=1}^{C} \|z_c^{\text{out}} - \text{sg}(z_c^{\text{in}})\|_2^2$.

The overall objective is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{PC}} + \lambda_2 \mathcal{L}_{\text{OV}} + \lambda_3 \mathcal{L}_{\text{RC}} + \lambda_4 \mathcal{L}_{\text{OC}}. \tag{1}$$

In all experiments, we set all $\lambda_i = 1$, assigning equal importance to each loss term. This configuration yielded stable optimization without additional hyperparameter tuning.

3.4 Region-based Organ-level Retrieval

We propose a region-based organ-level retrieval framework that operates on precomputed organ token representations, as illustrated in Figure 2. During database construction, all predefined regions are selected by the ROI Embedding Selector, and the trained ROA module generates organ token embeddings for each database image. For a database image d, we obtain an organ-wise representation set $Z_d = \{z_{d,c}^{\text{out}}\}_{c=1}^{C}, \quad z_{d,c}^{\text{out}} \in \mathbb{R}^D$. These representations are stored in a vector database and used for similarity computation. For a query image, the same feature extraction procedure is applied to obtain $Z_q = \{z_{q,c}^{\text{out}}\}_{c=1}^{C}$. When a region of interest (ROI) is specified, the ROI Embedding Selector extracts masked patch embeddings from the selected region, which are processed by the ROA module to produce organ-wise representations. To model organ presence, we introduce an Organ Visibility Recognition (OVR) network during training. The OVR network estimates the visibility probability of

each organ, $\hat{v}_c \in [0, 1]$, and these probabilities are used to construct a visibility-weighted aggregated representation:

$$\tilde{z}_x = \frac{1}{C} \sum_{c=1}^{C} \hat{v}_c z_{x,c}^{\text{out}}, \tag{2}$$

where $x \in \{q, d\}$. The visibility weights are predicted from the query image and applied to both the query and database representations.

The similarity between a query image I_q and a database image I_d is computed as the cosine similarity between their corresponding visibility-weighted representations, $\tilde{z}_q$ and $\tilde{z}_d$. Database images are ranked in descending order of this similarity score. Compared with global embedding-based retrieval, this query-driven weighting scheme focuses on anatomically observable organs and supports region-specific retrieval.

4 Experiments

4.1 Datasets

We construct the training and evaluation datasets using the TotalSegmentator (TS) dataset (Wasserthal et al., 2023, Version 2). The dataset consists of 1,228 volumes and 317,863 slices in total. The data split follows the official train/test partition defined in Vista3D (26). The training set includes 980 volumes (252,468 slices), and the test set comprises 248 volumes (65,395 slices). For retrieval evaluation, we build a slice-level dataset. The database set is created by uniformly sampling slices from the training volumes at intervals of 10 slices, in order to control slice density and reduce redundancy. This results in 25,696 database slices. The query set is constructed by randomly selecting five slices from each test volume, yielding a total of 1,241 query slices. To analyze the effect of region configuration on retrieval performance, we define four query region types: (1) a single randomly selected region, (2) two randomly selected regions, (3) three randomly selected regions, and (4) whole-image queries. All query types use the same number of queries (1,241) and the same database set (25,696 slices), enabling comparison under consistent evaluation conditions.

4.2 Evaluation metric

We adopt a region-centric evaluation protocol inspired by Jush et al. (27), which evaluates retrieval performance for each anatomical region independently. Given our slice-level retrieval setting, we adopt organ-level Precision@K to incorporate ranking information. This metric measures the proportion of retrieved slices within the top-K results that contain the target organ. For a given organ i, Precision$_i$@K is defined as

$$\text{Precision}_i@K = \frac{1}{|Q_i|} \sum_{q \in Q_i} \frac{|\{s \in R_q@K \mid i \in \mathcal{L}(s)\}|}{K}, \tag{3}$$

where Q_i denotes the set of query slices containing organ i, $R_q@K$ represents the set of top-K retrieved slices for query q, s denotes a retrieved slice, and $\mathcal{L}(s)$ is the set of organ labels annotated for slice s. The notation $|\cdot|$ indicates set cardinality. Note that Precision$_i$@K is not necessarily monotonically increasing with respect to K.

The overall performance is computed by averaging across all N evaluated organs:

$$\text{Precision}@K = \frac{1}{N} \sum_{i=1}^{N} \text{Precision}_i@K. \tag{4}$$

4.3 Comparison with Baselines

Recent studies (9; 28) have shown that pretrained Transformer-based vision foundation models trained with self-supervised learning achieve competitive performance in medical image retrieval.

Table 1: Retrieval performance (P@K) under different region configurations.

Region Configuration	Method	P@1	P@3	P@5	P@10	P@20
1 Region	Random	0.169	0.162	0.161	0.156	0.160
	ViT	0.150	0.163	0.162	0.164	0.164
	DINOv3	0.158	0.150	0.148	0.151	0.149
	DreamSim	0.117	0.113	0.108	0.109	0.110
	Proposed Model	**0.943**	**0.942**	**0.944**	**0.945**	**0.942**
2 Regions	Random	0.169	0.161	0.161	0.158	0.159
	ViT	0.336	0.323	0.314	0.314	0.308
	DINOv3	0.307	0.305	0.303	0.298	0.295
	DreamSim	0.347	0.342	0.341	0.331	0.320
	Proposed Model	**0.935**	**0.931**	**0.932**	**0.930**	**0.928**
3 Regions	Random	0.165	0.161	0.160	0.160	0.158
	ViT	0.425	0.417	0.408	0.398	0.388
	DINOv3	0.413	0.406	0.398	0.391	0.380
	DreamSim	0.489	0.475	0.462	0.450	0.434
	Proposed Model	**0.954**	**0.953**	**0.951**	**0.947**	**0.943**
Whole Image	Random	0.165	0.158	0.159	0.158	0.160
	ViT	0.704	0.690	0.678	0.663	0.645
	DINOv3	0.715	0.699	0.688	0.672	0.653
	DreamSim	0.785	0.773	0.764	0.751	0.740
	Proposed Model	**0.891**	**0.885**	**0.881**	**0.874**	**0.865**

Motivated by these findings, we adopt the pretrained baselines used in Jush et al. (28), including Vision Transformer (ViT) (29), DINOv3 (30), and DreamSim (31). For ROI-based retrieval, the target organ is cropped using a bounding box, and region embeddings are extracted and compared with global database embeddings to compute similarity scores. We also include a random ranking baseline without learned features. This setup enables quantitative evaluation of the limitations of global-embedding-based retrieval and assessment of the proposed region-based strategy.

Table 1 reports retrieval performance under varying region configurations. The proposed model achieves the highest scores across all settings, outperforming ViT, DINOv3, DreamSim, and the random baseline across all configurations. The performance gap is particularly pronounced in single- and multi-region configurations, indicating that organ-wise representations are effective for region-focused retrieval. Although global embedding methods improve as more regions are selected, their performance remains lower than that of the proposed model. In the Whole Image setting, the performance gap decreases, yet our method still yields the best results. Overall, organ-level representation learning with visibility-weighted aggregation consistently enhances region-focused retrieval performance.

Figure 3 provides a qualitative comparison under different region configurations. In the Whole Image setting (Fig. 3(a)), retrieval is based on the entire slice, leading to selection driven by global abdominal appearance; consequently, localized structures such as the colon are not consistently preserved among top results. In contrast, the Single Region setting (Fig. 3(b)) uses the colon as the ROI, resulting in more consistent localization and morphology alignment with the query region. In contrast, baseline methods relying on simple bounding-box cropping exhibit less stable structural alignment due to limited contextual information. DINOv3 and DreamSim occasionally retrieve magnified regions, while ViT retrieves colon-containing slices, but with less consistent localization and morphological alignment compared to the proposed method.

4.4 Ablation Study

Table 2 presents the ablation results analyzing the contribution of each training component in the proposed model. The baseline configuration (RC+VR) achieves stable performance across region settings. When Patch-level Classification (PC) is introduced, consistent improvements are observed under all configurations. This indicates that patch-level supervision contributes to more spatially

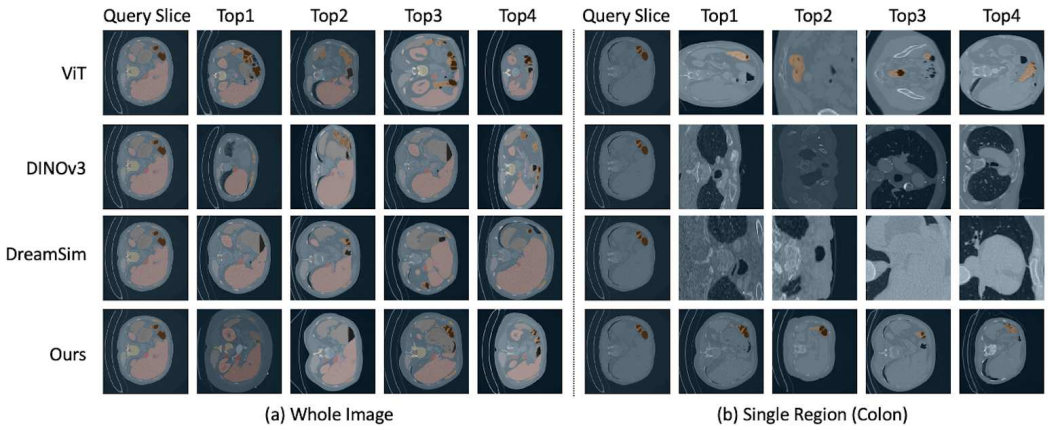

Figure 3: Qualitative comparison of retrieval results under different region configurations. The organ of interest in the query slice is highlighted with a colored segmentation mask. In the retrieved images, the corresponding organ is marked with the same color to facilitate visual comparison.

Table 2: Ablation results under different region configurations (Precision@K).

Region Configuration	Setting	P@1	P@3	P@5	P@10	P@20
1 Region	RC+VR	0.928	0.925	0.923	0.923	**0.983**
	RC+VR+PC	**0.946**	**0.947**	**0.948**	**0.946**	0.943
	RC+VR+PC+OC	0.943	0.942	0.944	0.945	0.942
2 Regions	RC+VR	0.915	0.909	0.904	0.902	0.898
	RC+VR+PC	**0.936**	0.929	0.928	0.924	0.921
	RC+VR+PC+OC	0.935	**0.931**	**0.932**	**0.930**	**0.928**
3 Regions	RC+VR	0.933	0.930	0.926	0.921	0.911
	RC+VR+PC	0.947	0.948	0.947	0.943	0.937
	RC+VR+PC+OC	**0.954**	**0.953**	**0.951**	**0.947**	**0.943**
Whole Image	RC+VR	0.876	0.869	0.863	0.854	0.845
	RC+VR+PC	0.889	0.885	**0.990**	0.874	**0.866**
	RC+VR+PC+OC	**0.891**	**0.885**	0.881	**0.874**	0.865

precise organ token representations, leading to improved region-based retrieval performance. With the addition of Organ Consistency (OC), the model achieves the highest performance under multi-region settings. In particular, under the 3 Region configuration, the model attains the highest P@1 score of 0.954. This result shows that enforcing consistency among organ token representations becomes increasingly beneficial when multiple organs are jointly considered. In contrast, under the Whole Image setting, the performance gains from OC are relatively modest, indicating that retrieval based on global input depends more heavily on the base representation learned by RC and VR. Overall, RC and VR establish the core representation, while PC and OC progressively refine it, resulting in incremental improvements in organ-centric retrieval performance. Although the relative contribution of each component may vary across region configurations, these variations reflect inherent differences between retrieval scenarios rather than instability in the model. Nevertheless, the full configuration (RC+VR+PC+OC) consistently achieves the best or near-best performance across all settings, indicating that a single unified configuration can be effectively adopted without query-dependent tuning.

4.5 Visualization and Interpretability Analysis

In this study, we adopt the pairwise similarity visualization method proposed by Black et al. (2022) (32) to analyze the spatial alignment between query and retrieved images. Figure 4 presents pairwise similarity maps between the query slice and the top-ranked retrieval results (top1–top5),

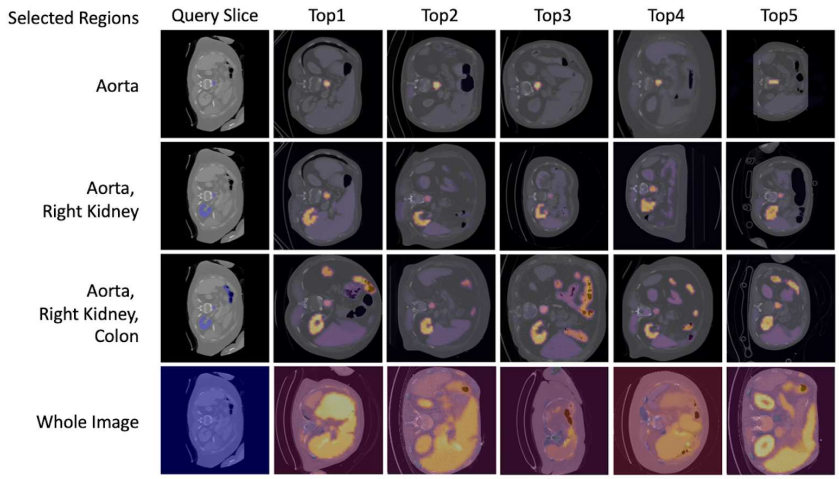

Figure 4: Pairwise rollout-based similarity maps under different region configurations. Each row shows the query slice and the top-ranked retrieval results (top1–top5). Regions contributing strongly to the similarity score are overlaid using a heatmap.

with each row corresponding to a different region configuration. When a single organ is selected as the query region, the similarity response is concentrated on the corresponding organ area in the retrieved images. When multiple organs are selected, the similarity maps jointly reflect the selected organs. As the number of selected regions increases, adjacent organ structures are progressively reflected in the similarity response. In contrast, under the whole image setting, retrieval is based on the entire slice, and similarity responses are primarily driven by global morphological resemblance. the highlighted regions are less consistent across retrieved images in terms of the areas referenced by the model, with relatively large organs such as the liver and kidneys showing stronger responses.

4.6 Limitations and Future works

This study demonstrates the potential of organ-level representation learning for region-focused medical image retrieval; however, several limitations remain. First, the proposed model is trained using precise organ segmentation annotations from the TotalSegmentator dataset, which assumes the availability of large-scale ground-truth segmentation labels. Second, the evaluation is primarily conducted at the slice level and thus does not fully capture organ relationships or spatial continuity at the 3D volume level. Third, comparative experiments are limited to ViT-based vision foundation models, and further validation across a broader range of retrieval models is necessary. Future work will include robustness evaluation under settings that rely on automatic segmentation models, such as Vista3D (26), and modeling organ relationships at the 3D volume level. In addition, we will analyze the impact of model parameters within the proposed architecture and perform systematic validation across diverse retrieval architectures to comprehensively evaluate the scalability and clinical applicability of the proposed approach.

4.7 Conclusions

In this study, we propose an organ-aware, region-centric representation learning framework to address the limitations of global similarity–based retrieval. The ROI Embedding Selector filters patch embeddings corresponding to the region of interest, while the Region-aware Organ Attention (ROA) module learns interactions between image patches and organ tokens to construct representations that precisely capture organ-specific structural and semantic information. In addition, during inference, we introduce a visibility-weighted aggregation strategy based on Organ Visibility Recognition, which prioritizes anatomically observable organs and enables retrieval focused on clinically meaningful regions of interest. By moving beyond simple global similarity computation and explicitly incorporating anatomical context, the proposed method supports more precise retrieval. This framework offers both methodological significance and practical potential for precise case retrieval and clinical decision support in real-world settings.

Acknowledgments

This work was supported by the Technology Innovation Program(or Industrial Strategic Technology Development Program-ATC+)(20023280, Development of life cycle management platform by providing artificial intelligence-based real-time model observability and explanability) funded By the Ministry of Trade, Industry and Resources(MOTIR, Korea)

References

[1] L. Cui and M. Liu, "An intelligent deep hash coding network for content-based medical image retrieval for healthcare applications," *Egyptian Informatics Journal*, vol. 27, p. 100499, 2024.

[2] J. Wang, H. Zhu, S.-H. Wang, and Y.-D. Zhang, "A review of deep learning on medical image analysis," *Mobile Networks and Applications*, vol. 26, no. 1, pp. 351–380, 2021.

[3] M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin, and M. P. Lungren, "Preparing medical imaging data for machine learning," *Radiology*, vol. 295, no. 1, pp. 4–15, 2020.

[4] J. Choe, H. J. Hwang, J. B. Seo, S. M. Lee, J. Yun, M. J. Kim, J. Jeong, Y. Lee, K. Jin, R. Park, J. Kim, H. Jeon, N. Kim, J. Yi, D. Yu, and B. Kim, "Content-based image retrieval by using deep learning for interstitial lung disease diagnosis with chest ct," *Radiology*, vol. 302, no. 1, pp. 187–197, 2022.

[5] Y. Nan, H. Zhou, X. Xing, G. Papanastasiou, L. Zhu, Z. Gao, A. F. Frangi, and G. Yang, "Revisiting medical image retrieval via knowledge consolidation," *Medical Image Analysis*, vol. 102, p. 103553, 2025.

[6] S. Agrawal, A. Chowdhary, S. Agarwala, V. Mayya, and S. Kamath S, "Content-based medical image retrieval system for lung diseases using deep cnns," *International Journal of Information Technology*, vol. 14, no. 7, pp. 3619–3627, 2022.

[7] R. Shetty, V. S. Bhat, and J. Pujari, "Content-based medical image retrieval using deep learning-based features and hybrid meta-heuristic optimization," *Biomedical Signal Processing and Control*, vol. 92, p. 106069, 2024.

[8] I. Issaoui, M. A. Alohali, W. Mtouaa, F. A. Alotaibi, A. Mahmud, and M. Assiri, "Archimedes optimization algorithm with deep learning assisted content-based image retrieval in healthcare sector," *IEEE Access*, vol. 12, pp. 29768–29777, 2024.

[9] A. Mahbod, N. Saeidi, S. Hatamikia, and R. Woitek, "Evaluating pre-trained convolutional neural networks and foundation models as feature extractors for content-based medical image retrieval," *Engineering Applications of Artificial Intelligence*, vol. 150, p. 110571, 2025.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[11] P. Das and A. Neelima, "An overview of approaches for content-based medical image retrieval," *International Journal of Multimedia Information Retrieval*, vol. 6, no. 4, pp. 271–280, 2017.

[12] A. Kumar, J. Kim, W. Cai, M. Fulham, and D. Feng, "Content-based medical image retrieval: A survey of applications to multidimensional and multimodality data," *Journal of Digital Imaging*, vol. 26, no. 6, pp. 1025–1039, 2013.

[13] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2015.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[17] C. M. Lo and C. Y. Hsieh, "Large-scale hierarchical medical image retrieval based on a multilevel convolutional neural network," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 9, no. 4, pp. 2782–2792, 2025.

[18] A. S. Susmitha and V. P. Namboodiri, "Analysis of transformers for medical image retrieval," 2024.

[19] S. Denner, D. Zimmerer, D. Bounias, M. Bujotzek, S. Xiao, R. Stock, L. Kausch, P. Schader, T. Penzkofer, P. F. Jäger, and K. Maier-Hein, "Leveraging foundation models for content-based image retrieval in radiology," *Computers in Biology and Medicine*, vol. 196, p. 110640, 2025.

[20] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," 2020.

[21] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, "Perceiver: General perception with iterative attention," in *The 38th International Conference on Machine Learning (ICML)*, pp. 4651–4664, PMLR, 2021.

[22] A. Jaegle, S. Borgeaud, J. B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer, O. J. Henaff, M. Botvinick, A. Zisserman, O. Vinyals, and J. Carreira, "Perceiver io: A general architecture for structured inputs and outputs," *The 10th International Conference on Learning Representations (ICLR)*, 2022.

[23] T. Yang, Y. Wang, Y. Lu, and N. Zheng, "Visual concepts tokenization," in *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, p. Article 2289, Curran Associates Inc., 2022.

[24] A. Ali-bey, B. Chaib-draa, and P. Giguère, "Boq: A place is worth a bag of learnable queries," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17794–17803, 2024.

[25] S. Song, S. Yoon, P. Jin, S. Kim, M. Tivnan, Y. Oh, R. Meng, L. Chen, Z. Lyu, and D. Wu, "Owt: A foundational organ-wise tokenization framework for medical imaging," *arXiv preprint arXiv:2505.04899*, 2025.

[26] Y. He, P. Guo, Y. Tang, A. Myronenko, V. Nath, Z. Xu, D. Yang, C. Zhao, B. Simon, and M. Belue, "Vista3d: A unified segmentation foundation model for 3d medical imaging," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20863–20873, 2024.

[27] F. Khun Jush, S. Vogler, and M. Lenga, "Content-based 3d image retrieval and a colbert-inspired re-ranking for tumor flagging and staging," *Journal of Imaging Informatics in Medicine*, 2025.

[28] F. K. Jush, S. Vogler, T. Truong, and M. Lenga, "Content-based image retrieval for multi-class volumetric radiology images: A benchmark study," *IEEE Access*, 2025.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[30] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, *et al.*, "Dinov3," *arXiv preprint arXiv:2508.10104*, 2025.

[31] S. Fu, N. Y. Tamir, S. Sundaram, L. Chai, R. Zhang, T. Dekel, and P. Isola, "Dreamsim: learning new dimensions of human visual similarity using synthetic data," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, (Red Hook, NY, USA), Curran Associates Inc., 2023.

[32] S. Black, A. Stylianou, R. Pless, and R. Souvenir, "Visualizing paired image similarity in transformer networks," in *The IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1534–1543, 2022.