Mutually Promoted Medical Image Segmentation and Classification via Prompt-Driven Interaction

Zijian Deng[1], Xiaoli Yao[1], Qijun Zhao[1]*, Iqra Shahzadi[1], and Hui Yang[2]*

[1] School of Computing Science, Sichuan University, Chengdu, Sichuan, China
qjzhao@scu.edu.cn, yh8806@163.com
[2] Department of Otolaryngology-Head and Neck Surgery, West China Hospital, Sichuan University, Chengdu, Sichuan, China

Abstract. The fine-tuned Segment Anything Model has shown strong potential for medical image segmentation, yet it remains class agnostic and cannot provide reliable categorical information alongside the predicted mask. This limits interpretability and increases the risk of selecting anatomically implausible regions when boundaries are ambiguous. Meanwhile, existing medical frameworks often treat segmentation and classification as separate processes, which weakens feature interaction and reduces robustness across diverse organs, lesions, and imaging modalities. We present BiMI, a mutually promoted framework that integrates a fine-tuned Segment Anything Model with a hierarchical classification branch. To regularize fine-grained recognition under data scarcity, BiMI introduces a Coarse2Fine Matrix that models hierarchical dependencies and enforces semantic consistency between coarse and fine predictions. To strengthen feature interaction between recognition and segmentation, BiMI further adopts a Global2Local Feature Enhancement module that aligns global context with local spatial cues. In addition, BiMI incorporates a Text Guidance module that expands the predicted coarse and fine labels into a compact structured description using a frozen language model, and encodes it into a semantic embedding with a frozen text encoder. The embedding conditions structure-aware candidate selection to rank multi-mask hypotheses within the box prompt, improving anatomical plausibility and segmentation stability. Experiments on six public medical datasets covering CT, MRI, dermoscopy, and endoscopy demonstrate that BiMI consistently improves hierarchical classification and achieves competitive segmentation performance against strong U-Net and SAM-based baselines, with clear gains on several challenging benchmarks.

Keywords: Promptable medical segmentation · Hierarchical classification· Multimodal learning· Feature fusion.

1 Introduction

Segmentation is a fundamental task in medical image analysis, aiming to delineate regions of interest such as organs, tissues, and lesions [8]. Accurate segmentation supports diagnosis, treatment planning, and longitudinal monitoring, while automated methods

* Corresponding authors.

reduce manual workload and improve consistency for large-scale analysis. Driven by advances in deep learning, modern segmentation models have achieved strong performance [9, 10], yet they are often optimized for specific targets and may generalize poorly across organs, lesions, and imaging modalities. Recently, promptable foundation models such as the Segment Anything Model [2] have shown promising transferability for medical segmentation [13]. However, the class-agnostic design of SAM provides masks without reliable categorical identity and offers limited interpretable mechanisms to verify anatomical plausibility, which restricts its clinical adoption [1].

In medical imaging workflows, segmentation is rarely used in isolation. Beyond delineating a region of interest, clinicians often require categorical identity to support reporting, retrieval, and follow-up comparison. A sequential pipeline that segments first and classifies later is sensitive to error propagation. Imprecise boundaries introduce background confounders and degrade fine-grained recognition, while ambiguous local appearance may lead a class-agnostic segmenter to select anatomically implausible regions when multiple mask hypotheses are possible within the same prompt.

Recent medical pipelines therefore perform classification on segmented regions, yet segmentation and classification are still commonly treated as separate stages with limited feature interaction. This design overlooks a practical coupling in medical images. Semantic identity provides an anatomical prior that helps reject false positives and stabilize ambiguous boundaries, and a reliable mask reduces irrelevant context for fine-grained recognition. Moreover, many existing designs are tailored to a narrow set of targets, which limits robustness across organs, lesions, and imaging modalities [17–19].

To address these limitations, we propose BiMI, a mutually promoted framework that integrates promptable segmentation with hierarchical classification to enforce semantic anatomical consistency. BiMI localizes the region of interest using a box prompt and performs hierarchical classification on the cropped region. A coarse-to-fine matrix named C2FM explicitly models the relationship between coarse and fine categories and encourages consistent predictions across label levels. Since local crops may be insufficient in complex anatomy, BiMI further introduces Global2Local feature enhancement named GFE, which aligns global features from the segmentation encoder with local representations to preserve global context and long-range dependencies.

BiMI also uses category information to stabilize mask selection within the localized region. The box prompt constrains the spatial search, but the segmentation decoder still produces multiple candidate masks that differ in scale, connectivity, and boundary leakage. We generate a compact semantic prior from the predicted coarse and fine labels. A frozen language model expands the label pair into a short structured description that emphasizes shape regularity, boundary characteristics, relative appearance, texture, and connectivity. A frozen text encoder converts the description into an embedding. Instead of injecting text into the segmentation decoder, we use the embedding to rank the candidate masks produced within the same region. For each candidate mask, we compute modality agnostic structural descriptors from its geometry, topology, and relative local appearance, and we match them with text conditioned structural weights to select the anatomically plausible mask using the Structure-Aware Candidate Selection block (SACS).

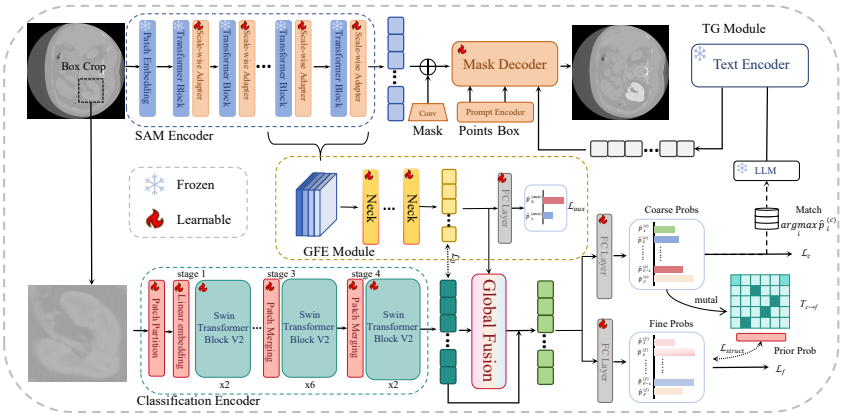

Fig. 1: Overview of BiMI, the input consists of a batch of image–mask–text pairs. **BiMI** establishes a bi-directional calibration mechanism: the segmentation mask defines the spatial attention for classification, while the classification branch provides semantic priors to rank the multi-mask hypotheses within the box prompt and to select anatomically plausible masks.

In summary, our main contributions are: (1) We propose BiMI, a mutually promoted framework bridging promptable segmentation and hierarchical classification; (2) We introduce C2FM and GFE to enforce hierarchical semantic consistency and align global-local features; (3) We design SACS, a differentiable text-guided module that ranks SAM's multi-mask hypotheses using explicit anatomical descriptors.

2 Method

As illustrated in Fig. 1, BiMI integrates promptable segmentation with hierarchical classification through three synergistic components. First, the Global2Local Feature Enhancement (GFE) module aligns global context from the SAM encoder with local features to stabilize classification. Simultaneously, the classification branch employs a Coarse2Fine Matrix (C2FM) to enforce hierarchical semantic consistency. Finally, the predicted labels condition the Structure-Aware Candidate Selection (SACS) to resolve segmentation ambiguity within the box prompt.

2.1 Hierarchical Classification with Coarse-to-Fine Guidance

Accurate fine-grained classification is often hindered by high intra-class variability and data scarcity in medical imaging. To mitigate this, we propose a soft hierarchical guidance strategy that leverages coarse anatomical priors to regularize fine-grained predictions. Specifically, we define a two-level taxonomy where each fine-grained category (specific target type) belongs to exactly one coarse category. Visual features extracted

by the SwinV2 backbone [7] are processed by two parallel heads to predict coarse probability $\hat{p}^{(c)}$ and fine probability $\hat{p}^{(f)}$, both supervised by standard cross-entropy losses.

Instead of treating the two heads independently, we explicitly model their dependency via a transition matrix. We construct a static matrix $T \in \mathbb{R}^{C_c \times C_f}$ derived from the co-occurrence statistics of the training set. To address the long-tailed distribution where common classes dominate, we compute the transition weights using inverse logarithmic frequency (specifically, $T_{i,j} \propto 1/\log(N_{i,j}+1)$, where $N_{i,j}$ is the frequency of fine class j in coarse category i), followed by row-wise normalization. This yields a balanced conditional probability $P(\text{fine}_j|\text{coarse}_i)$ that prevents the prior from overfitting to majority classes.

During training, we project the coarse prediction into the fine-grained label space to generate a prior distribution $\hat{p}^{(f)}_{\text{prior}} = \hat{p}^{(c)} \cdot T$. We then enforce semantic consistency by minimizing the Kullback-Leibler (KL) divergence between the model's fine-grained prediction and this coarse-guided prior:

$$\mathcal{L}_{\text{struct}} = \frac{1}{B} \sum_{i=1}^{B} \text{KL}\left(\hat{p}_i^{(f)} \parallel \hat{p}^{(f)}_{\text{prior},i}\right). \tag{1}$$

Unlike rigid logic rules, this soft regularization allows the network to learn subtle fine-grained features while penalizing predictions that violate the anatomical hierarchy (e.g., predicting a liver tumor in the kidney region), thus strictly limiting the search space for the fine-grained classifier.

2.2 Global2Local Feature Enhancement Module

Direct fusion of decoupled segmentation and classification features is prone to semantic misalignment. To address this, we propose the Global2Local Feature Enhancement (GFE) module to align global context with local ROI semantics.

Multi-scale Feature Aggregation. We first introduce Scale-wise Adapters (SA) into the frozen SAM encoder. Parallel dilated convolutions ($d = \{1, 2, 4\}$) augmented with channel attention are employed to capture multi-granular spatial context. The multi-scale outputs are summed and fused via a 1×1 convolution with a residual connection. We aggregate features from the last four SAM layers to form a context-enriched global representation $\mathbf{X}_g$.

To map $\mathbf{X}_g$ into the semantic space of local classification features $\mathbf{X}_c$, we employ a Feature Alignment (FA) mechanism. Inspired by SAM2CLIP, we utilize a transformer neck to project $\mathbf{X}_g$ alongside positional (PE) and level embeddings (LE):

$$\mathbf{X}'_g = \text{Proj}\left(\text{Neck}\left(\mathbf{X}_g + \text{PE} + \text{LE}\right)\right). \tag{2}$$

To enforce explicit distributional consistency between the global and local branches, we apply an MSE-based distillation loss:

$$\mathcal{L}_{\text{align}} = \frac{1}{B} \sum_{i=1}^{B} \left\| \mathbf{X}'_{g,i} - \mathbf{X}_{c,i} \right\|_2^2. \tag{3}$$

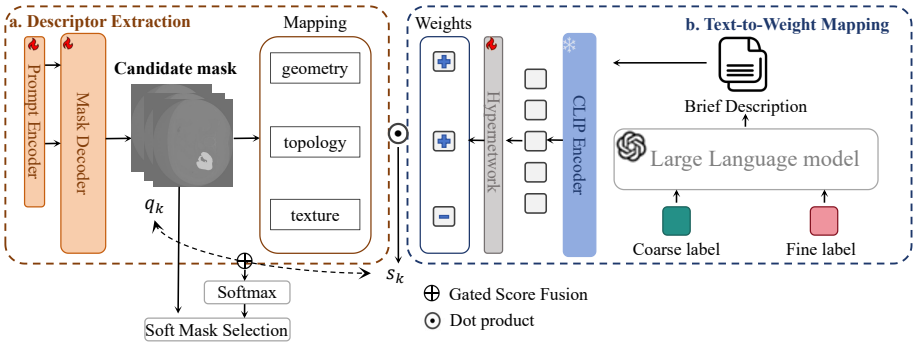

Fig. 2: Overview of Text-Guided module.

This alignment step is essential because GFE is not designed as direct global-feature injection; without FA, unaligned global context may introduce semantic noise rather than useful complementary cues. Furthermore, we introduce an auxiliary binary classification head (lesion vs. non-lesion) on $\mathbf{X}'_g$ to preserve global discriminative cues and mitigate task conflicts. This is supervised by a cross-entropy loss denoted as L_{aux}. Finally, the aligned global features $\mathbf{X}'_g$ and local features $\mathbf{X}_c$ are fused via a cross-attention mechanism, followed by a residual connection to yield $\mathbf{X}_{\text{final}}$ for hierarchical classification.

2.3 Promptable Segmentation with Text Guidance

While box prompts provide localization, they lack semantic awareness to resolve in-region ambiguity. To address this, we introduce Structure-Aware Candidate Selection (SACS), which converts semantic predictions into explicit anatomical constraints to rank SAM's hypotheses.

As shown in Fig. 2, we first identify the target semantics by selecting the most confident category pair from the classification branch: $c^* = \arg\max \hat{p}^{(c)}$ and $f^* = \arg\max \hat{p}^{(f)}$. A frozen LLM expands this pair into a structured description, encoded by CLIP [3] into a semantic embedding $\mathbf{e}$. Instead of generic feature matching, we use $\mathbf{e}$ to generate a dynamic anatomical prior via a lightweight hypernetwork that projects $\mathbf{e}$ into a structural weight vector $\mathbf{w}$:

$$\mathbf{w} = \mathbf{W}_2 \text{ReLU}(\mathbf{W}_1 \mathbf{e}). \tag{4}$$

The projection layer operates without a final activation function, enabling the network to learn positive weights that reward anatomical consistency and negative weights that penalize structural violations.

For each of the K candidate masks from SAM, we extract a descriptor vector D_k covering three modality-agnostic dimensions: **geometry** (compactness, solidity), **topology** (connectivity), and **texture** (gradient entropy). The anatomical compatibility score

Table 1: Performance comparison (DSC / IoU) on six medical segmentation datasets.

Models	AbdomenCT1K		ISIC2018		Kvasir-SEG		BraTS2021		KiTS19		BTCV	
	DSC	IoU	DSC	IoU	DSC	IoU	DSC	IoU	DSC	IoU	DSC	IoU
U-Net-based methods												
U-Net	81.6	70.3	88.3	79.7	86.5	76.7	85.1	74.2	89.0	81.2	83.6	71.8
Swin UNETR	87.1	77.2	89.4	82.2	87.5	78.7	88.4	81.8	90.9	83.3	86.9	78.1
ResUNet	84.6	73.3	87.1	78.2	89.2	81.5	78.4	71.3	89.2	82.4	-	-
nnUNet	91.3	**85.7**	90.0	78.5	81.8	74.6	88.5	80.6	83.8	73.9	80.2	66.9
TransUNet	88.5	80.1	89.4	82.2	88.7	79.2	86.6	79.0	89.5	82.0	83.8	72.1
SAM-based methods												
SAM	74.4	61.2	60.4	47.3	80.8	69.8	51.2	44.6	72.9	60.0	63.1	46.1
MedSAM	88.7	80.7	88.0	78.9	83.4	73.7	83.6	75.6	84.7	73.9	82.0	69.5
MSA	89.6	81.3	92.0	85.4	87.6	77.3	90.5	83.0	91.1	83.5	**89.8**	**81.5**
SAM-Med2D	90.5	83.6	92.2	86.2	87.9	77.9	**91.2**	**83.8**	91.5	85.1	86.7	79.6
LeSAM	-	-	92.7	86.8	88.5	79.3	-	-	91.9	85.0	-	-
BiMI (Ours)	**92.0**	82.4	**94.8**	**90.4**	**89.6**	**81.6**	90.2	82.3	**92.0**	**85.4**	88.2	79.4

is computed as $s_k = \sigma(\mathbf{w}^\top D_k)$. To balance geometric quality with semantic consistency, we fuse s_k with SAM's IoU prediction q_k via a text-adaptive gate $\alpha = \sigma(\mathbf{u}^\top \mathbf{e})$:

$$F_k = \alpha q_k + (1 - \alpha) s_k. \tag{5}$$

To enable end-to-end training, we formulate the selection as a differentiable soft-ranking process. The fused scores are normalized into selection probabilities $\pi_k = \mathrm{Softmax}(F_k/\tau)$, and the final mask $\hat{y}$ is computed as a weighted combination. The selection mechanism is supervised by aligning the predicted ranking with the ground-truth IoU distribution using KL divergence:

$$\mathcal{L}_{\text{sacs}} = \mathrm{KL}\left(\mathrm{Softmax}(\mathrm{IoU}(M_k, y)/\tau) \parallel \pi_k\right). \tag{6}$$

The segmentation backbone is supervised by a combination of Focal loss, Dice loss, and an IoU regression term:

$$\mathcal{L}_{\text{seg}} = \lambda_f \cdot \mathcal{L}_{\text{Focal}}(\hat{y}, y) + \mathcal{L}_{\text{Dice}}(\hat{y}, y) + \lambda_{\text{I}} \cdot \left\| \hat{\text{IoU}} - \text{IoU}(\hat{y}, y) \right\|_2^2. \tag{7}$$

Finally, the overall objective integrates the hierarchical classification, segmentation, and ranking terms:

$$L_{\text{total}} = \alpha_0 L_{\text{coarse}} + \delta L_{\text{fine}} + \beta L_{\text{seg}} + \gamma(L_{\text{align}} + L_{\text{aux}}) + \omega L_{\text{struct}} + \eta L_{\text{sacs}}. \tag{8}$$

3 Experiments

3.1 Datasets

We construct a large-scale medical image segmentation-classification dataset for model training, leveraging 1,031,240 images, covering 14 coarse-grained and 78 fine-grained categories.

To comprehensively evaluate the performance of our approach, we conducted experiments on multiple medical imaging datasets. On the AbdomenCT1K [20] and BTCV [21], we evaluated the model's performance in multi-organ segmentation, while its adaptability to lesion segmentation was tested on ISIC2018 [24], Kvasir-SEG [25], BraTS2021 [26], KiTS19 [22] covering various imaging modalities and lesion types. Furthermore, these test datasets are merged into a unified dataset to verify the model's category classification performance. All datasets used in this study are publicly available benchmarks. We followed the corresponding dataset terms of use and cite the original dataset publications. No new patient data were collected in this study.

3.2 Implementation details

Models are implemented in PyTorch and trained on an NVIDIA A100 GPU. 3D volumes are processed slice-by-slice along the axial plane, with images and masks resized to 256×256. For classification, ROIs are extracted via box prompts and resized to 224×224. We optimize the network using Adam (initial LR=1×10^{-4}, decayed by 0.5 at epochs 5 and 10) with a batch size of 100.

We employ a two-stage training strategy. In the first stage, we precompute a text embedding bank for all coarse and fine label pairs by running the frozen LLM once and encoding its descriptions with the frozen text encoder. The segmentation model is then trained for 10 epochs by retrieving the embedding indexed by the ground-truth label pair, updating only the adapter, the decoder, and the SACS scoring head. In the second stage, the classification branch is trained independently for 50 epochs without C2FM to avoid early interference from coarse-level predictions. Finally, after integrating both branches and training another 10 epochs, we freeze the segmentation branch and continue training the classification branch for another 40 epochs. This staged schedule avoids unstable early coupling, as unreliable early predictions may introduce noisy priors and mislead SACS-based candidate selection.

The loss function combines Focal Loss and Dice Loss with a weighting ratio of 20:1. The loss weights α_0, β, δ, ω, γ, and η are set to 0.8, 1, 1.2, 0.8, 0.5, and 0.3, respectively.

3.3 Comparative Experiment Results

We compare our proposed method BiMI with a series of strong baselines on both segmentation and classification tasks [9, 10, 12, 8, 11, 2, 13, 16, 15, 14]. As shown in Table 1, BiMI achieves state-of-the-art segmentation results on ISIC2018, Kvasir-SEG and KiTS19, with Dice scores reaching 94.8%, 89.6% and 92.0%, outperforming previous methods. On other datasets such as BraTS2021, AbdomenCT1K, and BTCV, BiMI consistently ranks among the top, demonstrating robust generalization across various lesion and organ segmentation scenarios.

For classification, Table 2 presents coarse and fine-grained accuracy across six datasets. BiMI surpasses both cropped baselines [5–7, 3, 4] (e.g., ResNet-101, Swin-ViT), achieving the best performance on four out of six datasets. Notably, BiMI achieves 53.6%/43.8% (coarse/fine) on AbdomenCT1K and 57.2%/45.1% on BTCV, significantly ahead of existing methods. These results verify that BiMI effectively combines

Table 2: Comparison of coarse/fine classification accuracy across datasets.

Models	AbdomenCT1K		ISIC2018		Kvasir-SEG		BraTS2021		KiTS19		BTCV	
	Coarse	Fine	Coarse	Fine	Coarse	Fine	Coarse	Fine	Coarse	Fine	Coarse	Fine
ResNet-50	42.5	22.5	92.8	91.8	84.2	76.4	90.4	32.3	33.7	30.0	41.6	33.3
ResNet-101	45.3	31.9	93.9	91.0	85.5	79.6	92.5	**38.6**	36.2	29.7	46.1	40.3
ViT	43.0	35.8	98.5	95.1	89.3	84.0	91.7	33.5	37.0	32.2	56.9	42.9
Swin-ViT v2	44.3	40.4	97.7	97.0	**92.8**	**92.6**	92.4	34.1	38.6	33.0	55.5	41.0
CLIP	17.6	6.4	27.3	0.7	16.9	–	23.1	19.5	29.3	1.2	15.4	3.2
+Adapter	35.5	15.5	96.9	96.8	90.4	89.9	84.7	32.7	35.9	20.6	35.6	17.5
BioMedCLIP	28.4	14.1	95.1	93.2	75.0	64.8	72.6	26.9	36.2	24.9	42.3	36.9
BMC-CLIP	34.6	17.1	95.4	91.1	86.9	72.6	83.1	21.1	35.7	21.4	39.8	41.3
BiMI (Ours)	**53.6**	**43.8**	**99.0**	**97.7**	89.2	86.0	**92.9**	34.4	**41.1**	**34.9**	**57.2**	**45.1**

Table 3: Ablation study of each module on total test set.

Adapter	SA	TG/SACS	C2FM	FA	GFE	Seg.		Cls.	
						IoU	DSC	Coarse	Fine
×	×	×	×	×	×	60.0	72.4	79.9	50.8
✓	×	×	×	×	×	77.6	85.8	80.1	52.3
×	✓	×	×	×	×	78.0	86.1	80.0	52.0
×	✓	✓	×	×	×	78.5	86.9	80.0	52.1
×	✓	✓	✓	×	×	78.8	87.3	81.8	52.5
×	✓	✓	✓	×	✓	73.9	85.3	73.4	45.3
×	✓	✓	✓	✓	✓	**79.1**	**87.7**	**82.4**	**53.2**

segmentation and classification, capturing both local details and global semantics, and sets a strong benchmark for integrated medical image segmentation and classification.

3.4 Ablation Experiment Results

We lastly assess the contribution of the main components through ablation (Table 3). Starting from a baseline without any modules, we then add Adapter, SA, TG/SACS, and C2FM sequentially, each yielding consistent improvements in segmentation and classification. Notably, adding GFE without FA decreases performance, indicating that direct injection of unaligned global features can introduce semantic mismatch. In contrast, the complete GFE+FA design achieves the best overall performance, confirming that FA is necessary for making global context beneficial. The full model with all modules achieves the best results, showing that each component plays a positive role and that their joint design is essential for robust segmentation and classification.

4 Conclusion and Future Work

In this paper, we presented BiMI, a mutually promoted framework that endows SAM with semantic awareness and hierarchical anatomical constraints. By integrating GFE

and SACS, BiMI effectively bridges the gap between class-agnostic segmentation and fine-grained medical recognition. Experiments across six diverse datasets demonstrate that BiMI not only provides reliable diagnostic interpretability but also utilizes semantic priors to significantly improve segmentation robustness, reducing the risk of hallucinating lesions in ambiguous regions. Currently, BiMI relies on manual box prompts. In future work, we aim to develop adaptive bounding box refinement strategies to further automate the clinical workflow and accommodate variable prompt inaccuracies.

Acknowledgments. This study was supported by the Chengdu Municipal Science and Technology Bureau Key R&D Support Program Technology Innovation R&D Project under Grant No. 2026-YF08-00338-GX.

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

References

1. Huang, Y., and Yang, X., and Liu, L.: Segment Anything Model for Medical Images? Medical Image Analysis, (2024)
2. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., et al.: Segment Anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
3. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning Transferable Visual Models From Natural Language Supervision. In: Proceedings of the International Conference on Machine Learning (ICML) (2021)
4. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: CLIP-Adapter: Better Vision-Language Models with Feature Adapters. International Journal of Computer Vision **132**, (2024)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image classification at scale. In: Proceedings of the International Conference on Learning Representations (ICLR) (2021)
7. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin Transformer V2: Scaling Up Capacity and Resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
8. Isensee, F., Jaeger, P.F., Kohl, S., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**, (2021)
9. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer (2015)
10. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: UNETR: Transformers for 3D medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2022)

11. 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)
12. Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C.: ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing **162**, (2020)
13. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment Anything in Medical Images. Nature Communications (2024)
14. Gu, Y., Wu, Q., Tang, H., Mai, X., Shu, H., Li, B., Chen, Y.: LeSAM: Adapt Segment Anything Model for Medical Lesion Segmentation. IEEE Journal of Biomedical and Health Informatics (2024)
15. Cheng, J., Ye, J., Deng, Z.: SAM-Med2D. arXiv:2308.16184 (2023)
16. Wu, J., Wang, Z., Hong, M., Ji, W., Fu, H., Xu, Y., Xu, M., Jin, Y.: Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation. Medical Image Analysis (2025)
17. Yuan, H., Li, X., Zhou, C., Li, Y., Chen, K., Loy, C.C.: Open-Vocabulary SAM: Segment and Recognize Twenty-thousand Classes Interactively. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024)
18. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. arXiv preprint arXiv:2401.14159 (2024)
19. Zhou, C., Loy, C.C., Dai, B.: Extract Free Dense Labels from CLIP. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022)
20. Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., et al.: AbdomenCT-1K: Is Abdominal Organ Segmentation a Solved Problem? IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
21. Landman, B., Xu, Z., Iglesias, J., Styner, M., Langerak, T., Klein, A.: MICCAI Multi-Atlas Labeling Beyond the Cranial Vault–Workshop and Challenge. In: Proceedings of the MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge (2015)
22. Heller, N., Sathianathen, N., Kalapara, A., Walczak, E. and Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., et al.: The KiTS19 Challenge Data: 300 Kidney Tumor Cases with Clinical Context, CT Semantic Segmentations, and Surgical Outcomes. arXiv preprint arXiv:1904.00445 (2019)
23. Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.-H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., et al.: CHAOS Challenge – Combined (CT-MR) Healthy Abdominal Organ Segmentation. arXiv preprint arXiv:2001.06535 (2020)
24. Ali, R., Hardie, R.C., De Silva, M.S., Kebede, T.M.: Skin Lesion Segmentation and Classification for ISIC 2018 by Combining Deep CNN and Handcrafted Features. arXiv preprint arXiv:1908.05730 (2019)
25. Jha, D.: Kvasir-SEG: A Segmented Polyp Dataset. In: MultiMedia Modeling. MMM 2020. Lecture Notes in Computer Science, vol. 11962. Springer (2020)
26. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. arXiv preprint arXiv:2107.02314 (2021)
27. Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al.: AMOS: A Large-Scale Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)