Unified Image-to-Image Generation for Diverse Medical Vision Tasks

***Abstract*—This paper introduces a novel framework that unifies various medical vision tasks, including synthesis, segmentation, denoising, and inpainting, into a single image-to-image generation process. By treating these tasks as conditional image generation problems, the proposed approach enables a generalist model to handle diverse inputs and outputs across different modalities and datasets. The effectiveness of this unification strategy is demonstrated through a comprehensive evaluation on a curated medical vision benchmark, showcasing its potential to simplify and enhance medical image analysis.**

***Index Terms*—Medical Image Analysis, Vision Transformer, In-Context Learning, Generalist Models, Image-to-Image Generation, Segmentation, Cross-Modal Synthesis, Inpainting, Denoising, Multi-Task Learning, Medical AI, Unified Benchmarking, Few-Shot Learning, Transferability and Multimodal Imaging**

I. Introduction

Accurate interpretation of medical imaging is a fundamental prerequisite for timely diagnosis and effective treatment of a wide array of health conditions [1], [2]. Recently, deep learning techniques have significantly advanced the field of medical image analysis by addressing various critical tasks, such as segmenting anatomical regions [3]–[5], identifying disease-specific locations [6]–[12], and synthesizing images across different modalities like MRI [13]–[16]. Despite their high task-specific performance, these models—commonly referred to as "specialist models"—are generally tailored for specific imaging modalities, anatomical regions, or tasks. Consequently, their effectiveness diminishes when applied to new tasks or when required to operate across domains with heterogeneous data.

To overcome these limitations, there has been a recent shift in focus towards building **generalist models** for medical AI [17], [18]. These models are designed for universal applicability across diverse medical imaging tasks using a single training cycle. Generalist frameworks typically achieve this by unifying the input and output space while leveraging prompts provided by users to dynamically specify tasks. Notable attempts such as MedSAM [19]–[21] have shown promising results in segmentation; however, they remain limited in scope, primarily supporting only segmentation-related tasks.

Inspired by the powerful in-context learning paradigms pioneered in NLP [22], [23] and their extensions to vision tasks [24]–[26], we introduce **Medical Vision Generalist (MVG)**, the first in-context learning-based generalist model tailored for the medical imaging domain. As shown in Figure 1, MVG is capable of performing multiple medical imaging tasks—including segmentation, cross-modal synthesis, inpainting, and denoising—across a range of modalities such as CT, MRI, X-ray, and micro-ultrasound.

MVG operates by converting all medical imaging tasks into a unified image-to-image generation framework. To standardize diverse input/output spaces, it applies an in-context coloring strategy, encoding all outputs into a single-channel format. This enables task execution based solely on visual prompts without requiring task-specific heads or model retraining. MVG further employs a hybrid learning strategy combining masked image modeling and auto-regressive learning to balance local and global context understanding. For inference, MVG selects contextually similar prompts to enhance task-specific predictions.

To rigorously evaluate MVG's capabilities, we develop a new benchmark incorporating 13 public datasets covering a wide spectrum of anatomical regions and imaging modalities. Experimental results reveal that MVG not only surpasses existing generalist models by significant margins (e.g., 0.735 mIoU on segmentation tasks, exceeding the previous best by 0.123 mIoU), but also demonstrates strong generalization with minimal labeled data. This positions MVG as a transformative step towards developing adaptable and scalable medical imaging solutions.

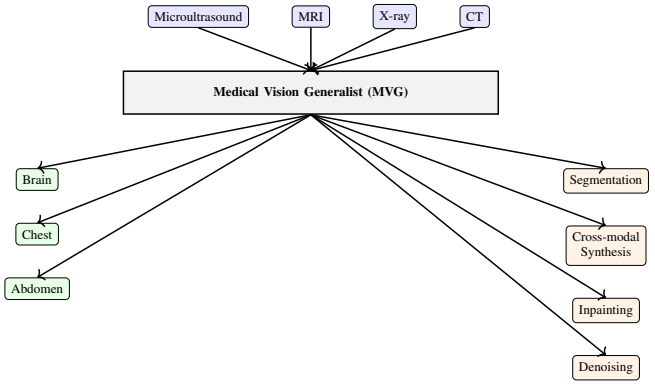

Fig. 1. Medical Vision Generalist supports four key imaging tasks across diverse modalities and anatomical regions.

II. Related Work

A. Medical Image Analysis

The domain of medical image analysis has witnessed substantial growth due to the integration of deep learning

methodologies, particularly in the area of medical image segmentation. A pioneering contribution in this space is the U-Net architecture [27], which introduced a robust encoder-decoder design with skip connections, enabling efficient localization and context utilization. Building upon this, nnU-Net [28] offered a self-configuring segmentation framework by incorporating a suite of engineering heuristics, thereby elevating segmentation performance without manual intervention. TransUNet [29] extended this paradigm by integrating the Vision Transformer (ViT) backbone to enhance feature representation, especially in scenarios demanding long-range spatial dependencies.

In pursuit of broader generalization across medical tasks, recent efforts have explored unified segmentation models. Noteworthy among these are "One Model to Rule Them All" MedSAM [19], and UniverSeg [21]. The latter adapts U-Net [27] to handle in-context learning by utilizing few-shot examples without requiring additional training on new segmentation tasks. Further, biomedGPT [20] emerges as a universal generative framework that bridges vision and language modalities within biomedical applications.

Unlike these models, our MVG introduces a novel in-context generation-based architecture capable of tackling not only segmentation but also a wider range of image-to-image translation tasks—including inpainting, denoising, and cross-modal synthesis—within a unified framework.

B. Universal Models and In-Context Learning

The emergence of large-scale transformer-based models has catalyzed the development of universal architectures capable of generalizing across tasks and modalities. A central advancement in this context is in-context learning (ICL), introduced by GPT-3 which enables models to infer tasks from a sequence of input-output pairs (prompts) without parameter updates. In language models, these prompts typically consist of a few text examples. Translating ICL into vision and multimodal domains remains an active area of research.

In the vision domain, Flamingo [45] was among the first to extend ICL with visual sequences and multimodal instructions. Perceiver IO [46] introduced a modular transformer capable of processing a broad range of data types including images, videos, and text. Other works such as AD [47] and DPT [48] have explored ICL within reinforcement learning, using distilled prompts or trajectories to inform decision-making policies.

MVG adopts this learning paradigm for medical vision by using a sequence of visual prompts consisting of paired images and labels. It unifies 13 distinct medical imaging tasks as a conditional generation problem, enabling few-shot task adaptation without re-training. By converting all tasks into a standard image format, MVG bridges task heterogeneity through a consistent, prompt-driven learning mechanism that is grounded in medical context.

III. Method

Traditional medical AI systems often require task-specific architectures and retraining when faced with new imaging modalities or tasks. In contrast, the proposed **Medical Vision Generalist (MVG)** aims to unify multiple medical imaging tasks—including segmentation, cross-modal synthesis, inpainting, and denoising—under a single image-to-image generation framework. MVG enables flexible and scalable task adaptation by leveraging in-context learning and conditional generation principles, without requiring task-specific output heads.

A. Task Formulation

MVG supports a variety of medical vision tasks, formulated as conditional image-to-image generation problems. The tasks investigated in this work include:

- **Segmentation**: This involves identifying anatomical structures (e.g., liver, kidney, prostate) in CT, MRI, X-ray, or micro-ultrasound scans. The output is a semantic mask that assigns class labels to image regions.
- **Cross-Modal Synthesis**: Given an image in one modality (e.g., CT), the model synthesizes its counterpart in a different modality (e.g., MRI) for the same anatomical region, supporting multimodal clinical workflows.
- **Inpainting**: For brain MRIs affected by gliomas, MVG reconstructs healthy tissue in tumor-affected regions [32], allowing integration with downstream tools like brain parcellation.
- **Denoising**: MVG reconstructs full-dose CT images from their low-dose counterparts, thereby supporting radiation-safe diagnostic imaging.

B. Unifying Input and Output Spaces

Let $x \in \mathbb{R}^{H \times W}$ denote an input image. MVG reformulates diverse tasks—each with unique output representations—into a unified image output format. This is achieved by mapping all outputs into a single-channel representation using **in-context coloring**, inspired by strategies in [25], [26]. Three coloring schemes are explored:

- **Binary Coloring**: Segmentation masks with N_k foreground classes are decomposed into N_k binary masks, each isolating one class. While precise, this approach requires multiple inferences.
- **Pre-defined Coloring**: Each class label is assigned a globally unique ID based on dataset and class index. Identical anatomical classes across datasets (e.g., "Liver") may receive different values.
- **Random Coloring**: Random colors are sampled for each class within a single task iteration. The same semantic label uses the same color in the prompt and task image. This encourages MVG to rely on spatial and contextual information rather than fixed color mappings.

Coloring is applied only to segmentation tasks, as other tasks involve continuous image outputs.

C. Task Unification via Conditional Image Generation

After normalizing the input and output spaces, all tasks are unified into a conditional generation format:

$$Y = \text{MVG}(X \mid P_x, P_y)$$

where X is the task image, P_x and P_y denote the prompt image and prompt label, and Y is the task label to be predicted.

MVG supports two conditional training paradigms:

1) Masked Image Modeling (MIM): In MIM, prompt and task images (and their labels) are concatenated to form a single 2x2 image grid. Random patches are masked, and the model reconstructs these regions:

$$p(x) = \prod_{i=1}^{M} p(x_i \mid x_{\notin x_M}, \theta) \tag{1}$$

where x_M represents masked patches and θ the model parameters.

MIM is well-suited for tasks focused on local detail restoration (e.g., inpainting, denoising), but it struggles with segmentation of small organs, which may be entirely masked and thus not visible in context.

2) Auto-Regressive Training: To preserve global spatial relationships, MVG also uses an auto-regressive learning paradigm. The input sequence is:

$$S = [P_{x_1}, P_{y_1}, ..., P_{x_n}, P_{y_n}, X, Y]$$

Each step predicts the next image in the sequence given all prior elements:

$$p(x) = \prod_{i=1}^{n+1} p(S_{2i} \mid S_1, ..., S_{2i-1}, \theta) \tag{2}$$

This strategy maintains holistic context, significantly improving segmentation quality.

3) Architecture: MVG uses a standard Vision Transformer (ViT) [30] as its backbone. The encoder consists of patch embeddings and transformer blocks. The decoder, following [31], employs two convolutional layers and aggregates four feature maps from different ViT layers to generate final predictions.

4) Inference Strategy: At inference, the model constructs the sequence $[P_x, P_y, X, \hat{Y}]$ where $\hat{Y}$ is the predicted label. To ensure contextual alignment, the prompt (P_x, P_y) is selected from training samples with spatial and anatomical properties most similar to the test image X. Specifically, given X_{TE} with N_{TE} slices and X_{TR} with N_{TR} slices, the prompt slice is:

$$\text{Prompt Index} = \left\lfloor \frac{n \cdot N_{TR}}{N_{TE}} \right\rfloor$$

This ensures structural consistency during in-context learning.

IV. Experiments

A. Implementation Details

1) Datasets: To evaluate the effectiveness of MVG across a wide variety of medical vision tasks, we compile a comprehensive benchmark of 13 publicly available datasets encompassing four major anatomical regions: **abdomen**, **pelvis**, **brain**, and **chest**. As outlined in Table I, the datasets span multiple imaging modalities including CT, MRI, X-ray, and micro-ultrasound. Collectively, this benchmark comprises 2.5 million training images.

TABLE I
DATASET OVERVIEW: MVG IS TRAINED ON 13 DATASETS SPANNING 4 BODY REGIONS AND 4 IMAGING MODALITIES.

Region	Dataset	Modality	Train	Test	Task
Abdomen	AMOS [3]	CT	240	120	Segmentation
	WORD [4]	CT	100	20	Segmentation
	BTCV [39]	CT	21	9	Segmentation
	AMOS [3]	MRI	60	50	Segmentation
Pelvis	MicroSegNet [40]	Micro-US	55	20	Segmentation
	PROMISE [40]	MRI	50	30	Segmentation
Brain	BraTS-GLI [**?**]	MRI	1251	219	X-Modal Synth.
	BraTS-Local [**?**]	MRI	1000	251	Inpainting
Chest	Low Dose [43]	CT	200	59	Denoising
	Defect [36]	X-ray	15	6	Segmentation
	ACDC [37]	MRI	100	50	Segmentation
	LA [38]	MRI	81	20	Segmentation
Whole	DeepLesion [42]	CT	25000	7120	Detection

2) Preprocessing: All CT scans are windowed to the range of $[-100, 200]$ for enhanced tissue contrast. Each image is resized to 512×512 and then randomly cropped to 448×448 as input. For out-of-distribution evaluation, we use the Medical Segmentation Decathlon (MSD) dataset [44] to test generalization capabilities.

3) Training Protocol: We use the AdamW optimizer with a weight decay of 0.05 and an initial learning rate of 1×10^{-3}, following a cosine annealing schedule with 5 warm-up epochs over 100 total training epochs. All experiments are conducted using 8 NVIDIA A5000 GPUs. We use 1 in-context prompt sample for both training and inference.

Segmentation tasks are sampled with a weight of 0.5, while the other tasks share the remaining 0.5. Data augmentation is limited to random cropping. Training employs 90% MIM and 10% auto-regressive training for non-segmentation tasks, and 100% auto-regressive training for segmentation.

4) Evaluation Metrics:

- **Segmentation**: Mean Intersection over Union (mIoU)
- **Cross-Modal Synthesis, Inpainting, Denoising**: Mean Absolute Error (MAE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM)

B. Generalist Performance Across 13 Medical Tasks

1) Baselines: We compare MVG against both specialist and generalist baselines:

- **Specialists**: ResNet-18 [34], UNet [27], VNet [35], TransUNet [29], nnUNet [28], and Pix2Pix [33] (for synthesis).
- **Generalists**: LVM [24], Painter [25], SegGPT [26], and UniverSeg [21].

2) Quantitative Results: As shown in Table II, MVG significantly outperforms other generalists across segmentation tasks, achieving an average mIoU of 0.79. Notably, it improves over SegGPT by 0.09 mIoU, Painter by 0.24, and LVM by 0.62 mIoU. Although specialist models like nnUNet achieve slightly higher scores on some datasets, they require separate models per task, unlike MVG.

TABLE II
SEGMENTATION MIOU COMPARISON WITH GENERALIST AND SPECIALIST MODELS

Method	AMOS CT	WORD	BTCV	AMOS MRI	MicroSegNet	PROMISE	Chest Defect	ACDC	LA
ResNet-18	0.55	0.50	0.51	0.53	0.67	0.75	0.62	0.69	0.68
UNet	0.81	0.83	0.82	0.81	0.90	0.91	0.89	0.86	0.83
VNet	0.70	0.75	0.72	0.73	0.90	0.89	0.86	0.87	0.84
TransUNet	0.80	0.82	0.84	0.82	0.94	0.90	0.88	0.88	0.84
nnUNet	0.87	0.90	0.91	0.88	0.97	0.93	0.90	0.90	0.89
UniverSeg	0.20	0.29	0.37	0.25	0.71	0.55	0.55	0.54	0.57
Painter	0.52	0.48	0.45	0.51	0.69	0.68	0.50	0.52	0.55
LVM	0.12	0.14	0.10	0.15	0.36	0.30	0.10	0.12	0.13
SegGPT	0.66	0.66	0.65	0.71	0.88	0.75	0.68	0.70	0.71
MVG (Ours)	**0.73**	**0.74**	**0.73**	**0.74**	**0.91**	**0.85**	**0.79**	**0.85**	**0.81**

3) Qualitative Results: Figure 2 visualizes segmentation predictions from MVG across various datasets. The outputs demonstrate MVG's robustness across anatomical regions and modalities.

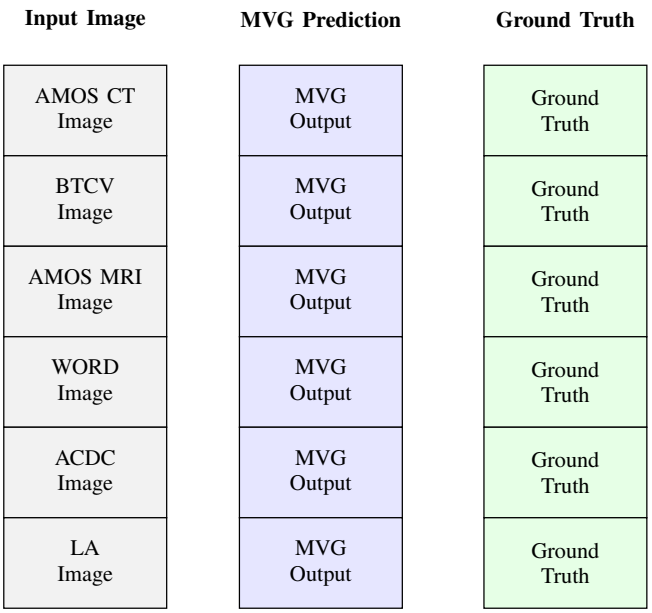

Fig. 2. Qualitative segmentation results across datasets. Each row shows an input image, the MVG model's prediction, and the ground truth segmentation.

C. Performance on Cross-Modal Synthesis, Inpainting, and Denoising

We further evaluate MVG's effectiveness on continuous output tasks—cross-modal synthesis, inpainting, and denoising—using metrics such as Mean Absolute Error (MAE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM). Table III compares MVG against both specialist and generalist models across these tasks.

MVG consistently outperforms Painter [25] and shows comparable performance to well-optimized specialist models. For example, in the cross-modal synthesis task, MVG yields a 0.002 lower MAE, a 0.69 higher PSNR, and a 0.009 improvement in SSIM compared to the best generalist baseline.

TABLE III
PERFORMANCE COMPARISON: CROSS-MODAL SYNTHESIS, INPAINTING, AND DENOISING

Method	**Synthesis**			**Inpainting**			**Denoising**		
	MAE	PSNR	SSIM	MAE	PSNR	SSIM	MAE	PSNR	SSIM
ResNet-18	0.026	20.98	0.860	0.008	30.98	0.959	0.022	30.52	0.709
Pix2Pix [33]	0.018	24.31	0.899	0.008	34.89	0.982	0.020	33.01	0.730
TransUNet	0.016	25.54	0.938	0.005	35.56	0.989	0.016	33.99	0.761
Painter	0.021	24.03	0.920	0.006	33.59	0.978	0.020	33.10	0.721
MVG (Ours)	**0.019**	**24.72**	**0.929**	**0.006**	**34.52**	**0.981**	**0.018**	**33.52**	**0.731**

D. MVG for Detection and Annotation-Format Adaptation

Although designed for generative tasks, MVG shows promise in more discriminative settings such as object detection. Using the DeepLesion [42] dataset, MVG overlays lesion bounding boxes onto output images. Unlike conventional object detectors, MVG encodes annotations directly as visual prompts, such as rectangles or circles, making it highly adaptable to diverse annotation formats.

Figure 3 illustrates MVG's performance on lesion localization, indicating strong potential for future multimodal annotation adaptation.

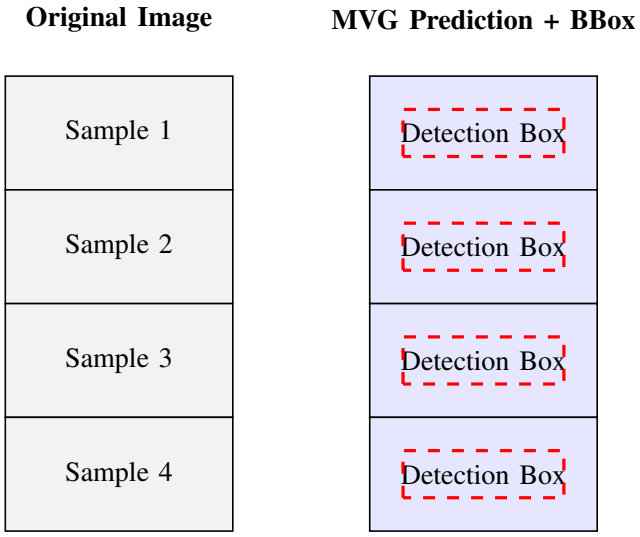

Fig. 3. Detection examples from the DeepLesion dataset. Left: original CT image. Right: MVG prediction with overlaid bounding box highlighting the lesion.

E. Data Efficiency and Generalization to Unseen Datasets

Thanks to its in-context learning capability, MVG requires minimal labeled examples to adapt to new datasets. We evaluate this on three unseen tasks from the Medical Segmentation Decathlon (MSD) [44].

TABLE IV
ZERO/FEW-SHOT GENERALIZATION TO MSD TASKS

Dataset	mIoU
MSD-Liver (prompt only)	0.84
MSD-Spleen (1-shot fine-tune)	0.87
MSD-Lung (1-shot fine-tune)	0.48

As shown in Table IV, MVG achieves competitive results on MSD-Liver without fine-tuning and improves further with just one training instance for MSD-Spleen and MSD-Lung.

F. Ablation Study

1) Color Encoding Strategies: To examine the impact of output representation schemes for segmentation, we compare binary, pre-defined, and random coloring methods. Table V reveals that random color assignments yield superior performance across nearly all datasets.

TABLE V
COMPARISON OF COLOR ENCODING STRATEGIES FOR SEGMENTATION

Method	AMOS CT	WORD	BTCV	AMOS MRI	MicroSegNet	PROMISE	Chest Def.	ACDC	LA
Binary	0.46	0.48	0.48	0.49	0.78	0.78	0.45	0.49	0.52
Pre-def.	0.60	0.62	0.62	0.61	0.89	0.86	0.71	0.74	0.73
Random	**0.73**	**0.74**	**0.73**	**0.74**	**0.91**	**0.85**	**0.79**	**0.85**	**0.81**

2) Training Strategy: Isolated vs. Unified: To assess the benefits of unified learning across datasets, we compare isolated per-dataset training against joint training. Results in Table VI demonstrate that unified training significantly enhances model performance, yielding up to 0.14 mIoU improvement over isolated models.

TABLE VI
ISOLATED VS. UNIFIED TRAINING STRATEGY

Setting	AMOS CT	WORD	BTCV	AMOS MRI	MicroSegNet	PROMISE	Chest Def.	ACDC	LA
Isolated	0.55	0.57	0.57	0.58	0.80	0.77	0.70	0.69	0.68
Unified	**0.73**	**0.74**	**0.73**	**0.74**	**0.91**	**0.85**	**0.79**	**0.85**	**0.81**

3) Masked Image Modeling vs. Auto-Regressive Training: We further compare the impact of training strategies. Table VII shows that auto-regressive learning drastically improves performance over masked image modeling (MIM), especially for small organ segmentation tasks where random masking removes essential context.

TABLE VII
IMPACT OF MASKED IMAGE MODELING VS. AUTO-REGRESSIVE TRAINING

Method	AMOS CT	WORD	BTCV	AMOS MRI	MicroSegNet	PROMISE	Chest Def.	ACDC	LA
MIM (50%)	0.56	0.48	0.46	0.54	0.70	0.66	0.50	0.50	0.52
MIM (75%)	0.53	0.42	0.44	0.52	0.70	0.63	0.48	0.50	0.51
Auto-Reg.	**0.73**	**0.74**	**0.73**	**0.74**	**0.91**	**0.85**	**0.79**	**0.85**	**0.81**

V. CONCLUSION

In this work, we introduced **Medical Vision Generalist (MVG)**, a unified and scalable framework that performs multiple medical imaging tasks—segmentation, cross-modal synthesis, inpainting, and denoising—within a single image-to-image generation pipeline. Unlike traditional specialist models that require separate architectures and retraining for each task or modality, MVG leverages an in-context learning paradigm that enables flexible task specification via visual prompts.

By unifying the input and output spaces using a shared visual representation and adopting a hybrid training scheme combining masked image modeling and auto-regressive learning, MVG effectively learns from contextual information. Our model eliminates the need for task-specific heads and demonstrates remarkable generalization capability across modalities and anatomical regions.

To assess MVG comprehensively, we curated the first large-scale benchmark for generalist medical vision models, encompassing 13 datasets across CT, MRI, X-ray, and micro-ultrasound modalities. Experimental results confirm that MVG achieves state-of-the-art performance across a range of medical vision tasks while significantly outperforming prior generalist frameworks. Moreover, MVG scales effectively with data volume, adapts quickly to new datasets with minimal labeled samples, and generalizes across diverse clinical imaging conditions.

We believe that MVG represents a pivotal step toward democratizing medical AI by reducing the barriers of model specialization and retraining. By releasing our benchmark and model code, we hope to foster further research in building general-purpose, accessible, and context-aware medical imaging systems that can support a wide range of real-world clinical needs.

REFERENCES

[1] J. Cheng *et al.*, "ResGANet: Residual group attention network for medical image classification and segmentation," *Medical Image Analysis*, vol. 76, p. 102313, 2022.

[2] J. De Fauw *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature Medicine*, vol. 24, no. 9, pp. 1342–1350, 2018.

[3] Y. Ji *et al.*, "AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation," in *NeurIPS*, 2022.

[4] X. Luo *et al.*, "WORD: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from CT image," *arXiv preprint arXiv:2111.02403*, 2021.

[5] Y. Fu *et al.*, "A review of deep learning based methods for medical image multi-organ segmentation," *Physica Medica*, vol. 85, pp. 107–122, 2021.

[6] W. Zhu *et al.*, "Anatomynet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy," *Medical Physics*, vol. 46, no. 2, pp. 576–589, 2019.

[7] C. Zhao *et al.*, "Improved lung nodule segmentation with adversarial training," *Computers in Biology and Medicine*, vol. 134, p. 104444, 2021.

[8] Y. Huo *et al.*, "Harvesting, detecting, and characterizing liver lesions from large-scale multi-phase CT data via deep dynamic texture learning," *arXiv preprint arXiv:2006.15691*, 2020.

[9] C. Cheng *et al.*, "A flexible 3D heterophase CT HCC detection algorithm for generalizable and practical screening," *Hepatology Communications*, 2022.

[10] D. Ardila *et al.*, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature Medicine*, vol. 25, no. 6, pp. 954–961, 2019.

[11] R. Kim *et al.*, "AI tool for assessment of indeterminate pulmonary nodules detected with CT," *Radiology*, p. 212182, 2022.

[12] N. Heller *et al.*, "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge," *Medical Image Analysis*, vol. 67, p. 101821, 2021.

[13] C. Xie *et al.*, "MRI modality synthesis using deep learning: A survey," *arXiv preprint arXiv:2305.07327*, 2023.

[14] H. Li *et al.*, "The BraTS challenge 2023: Brain MR image synthesis for tumor segmentation (BraSyn)," *arXiv preprint arXiv:2305.08992*, 2023.

[15] S. Dayarathna *et al.*, "Deep learning-based synthesis of MRI, CT, and PET: Review and analysis," *Medical Image Analysis*, p. 103046, 2023.

[16] Z. Zhu *et al.*, "Cross-modality brain image synthesis using transformer networks," *arXiv preprint arXiv:2306.00123*, 2023.

[17] M. Moor *et al.*, "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, no. 7956, pp. 259–265, 2023.

[18] T. Tu *et al.*, "Towards generalist biomedical AI," *arXiv preprint arXiv:2307.14334*, 2023.

[19] J. Ma *et al.*, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[20] T. Zhang *et al.*, "biomedGPT: A unified model for biomedical vision-language understanding and generation," *arXiv preprint arXiv:2303.12251*, 2023.

[21] V. I. Butoi *et al.*, "UniverSeg: Universal medical image segmentation," in *ICCV*, 2023.

[22] T. Brown *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[23] OpenAI, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[24] Y. Bai *et al.*, "Sequential modeling enables scalable learning for large vision models," *arXiv preprint arXiv:2312.00785*, 2023.

[25] P. Wang *et al.*, "Painter: Representation learning for vision tasks with auto-regressive generation," *arXiv preprint arXiv:2303.17580*, 2023.

[26] P. Wang *et al.*, "SegGPT: Segmenting everything in context," *arXiv preprint arXiv:2304.03284*, 2023.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[28] F. Isensee *et al.*, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[29] J. Chen *et al.*, "TransUNet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[30] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[31] Y. Li *et al.*, "Exploring plain vision transformer backbones for object detection," in *ECCV*, 2022.

[32] F. Kofler *et al.*, "BraTS 2023: Local synthesis of healthy brain tissue via inpainting," *arXiv preprint arXiv:2305.08992*, 2023.

[33] P. Isola *et al.*, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017.

[34] K. He *et al.*, "Deep residual learning for image recognition," in *CVPR*, 2016.

[35] F. Milletari, N. Navab, and S. A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *3DV*, 2016.

[36] S. Candemir and S. Antani, "A review on lung boundary detection in chest X-rays," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, pp. 563–576, 2019.

[37] O. Bernard *et al.*, "Deep learning techniques for automatic MRI cardiac multi-structure segmentation and diagnosis," *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[38] C. Chen *et al.*, "Multi-task learning for left atrial segmentation on GE-MRI," in *STACOM-MICCAI*, pp. 292–301, 2019.

[39] J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: A survey," *Medical Image Analysis*, vol. 24, no. 1, pp. 205–219, 2015.

[40] G. Litjens *et al.*, "Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge," *Medical Image Analysis*, vol. 18, no. 2, pp. 359–373, 2014.

[41] H. Jiang *et al.*, "MicroSegNet: A deep learning approach for prostate segmentation on micro-ultrasound images," *Comput. Med. Imaging Graph.*, vol. 112, p. 102326, 2024.

[42] K. Yan *et al.*, "DeepLesion: Automated mining of large-scale lesion annotations and universal lesion detection with deep learning," *J. Med. Imaging*, vol. 5, no. 3, p. 036501, 2018.

[43] C. McCollough *et al.*, "Low dose CT image and projection data (LDCT-and-projection-data)," *Medical Physics*, vol. 48, pp. 902–911, 2021.

[44] M. Antonelli *et al.*, "The Medical Segmentation Decathlon," *Nature Communications*, vol. 13, no. 1, p. 4128, 2022.

[45] J. Alayrac *et al.*, "Flamingo: A visual language model for few-shot learning," in *NeurIPS*, 2022.

[46] A. Jaegle *et al.*, "Perceiver IO: A general architecture for structured inputs and outputs," *arXiv preprint arXiv:2107.14795*, 2021.

[47] M. Laskin *et al.*, "In-context reinforcement learning with algorithm distillation," *arXiv preprint arXiv:2210.14215*, 2022.

[48] J. Lee *et al.*, "Supervised pretraining can learn in-context reinforcement learning," in *NeurIPS*, 2024.