

# Few-Shot Medical Image Segmentation with Meta-Learning: An Empirical Study on Organ Segmentation

Ismail EL Jamiy<sup>\*†</sup>, Amel Aissaoui<sup>\*</sup>, Lotfi Chaari<sup>\*</sup>,

<sup>\*</sup>IRIT, Toulouse University, CNRS, Toulouse INP, UT3  
Toulouse, France.

e-mail: firstname.lastname@irit.fr

<sup>†</sup>ENSAM Casablanca, University Hassan 2  
Casablanca, Morocco

**Abstract**—Few-shot learning has emerged as a promising approach for medical image segmentation, where annotated data are scarce for certain organs or pathologies. In this work, we explore the use of the Model-Agnostic Meta-Learning (MAML) framework to enable rapid adaptation to new organ segmentation tasks in limited data context using the BTCV multi-organ dataset. A 2D U-Net model is trained within the MAML paradigm, providing a meta-learned initialization that can be fine-tuned with only a few annotated examples of an unseen organ. Experimental results across 3-, 5- and 10-shots settings demonstrate that MAML consistently achieves faster and more effective adaptation than conventional training and organ-specific pretraining, especially in very low-shot regimes. These findings highlight the potential of optimization-based meta-learning for organ-specific medical image segmentation in data-limited scenarios.

**Keywords**—Medical imaging, deep learning, few-shot learning, MAML, organ segmentation.

## I. INTRODUCTION

Medical image segmentation plays a pivotal role in numerous clinical applications, including surgical planning, disease diagnosis, and radiotherapy treatment. With the recent advancements in deep learning, deep convolutional neural networks (CNN), particularly U-Net, have become the standard for medical image segmentation [1]. However, these models need extensive training on large annotated datasets for optimal performance, which are unavailable (rare pathologies or imaging protocols), costly or time-intensive since they require inputs from experts with extensive clinical experience. Indeed, their performance heavily decreases in scarce data context. Few-shot learning (FSL) addresses this challenge by enabling training models with only a small number of annotated examples. Within FSL, meta-learning, and in particular the Model-Agnostic Meta-Learning (MAML) framework, has shown strong potential by learning initializations that adapt quickly to unseen tasks [2]. While it is one of the most widely used approaches, its applicability to medical image segmentation remains underexplored [3]. In this paper, we investigate the application of MAML to medical image

segmentation, with a particular focus on adapting to novel organs using the BTCV multi-organ dataset. The paper reviews related work, presents the MAML framework, describes our experiments, and concludes with perspectives.

## II. RELATED WORKS

### A. Deep Learning for Medical Image Segmentation

The U-Net architecture [1] has become the standard for a wide range of medical segmentation tasks. Its symmetric encoder-decoder structure, complemented by skip connections that propagate fine-grained spatial information from the contracting to the expansive path, has proven exceptionally effective for precise pixel-level localization. Since its introduction, U-Net has inspired a vast family of extensions designed to address different challenges in medical imaging. For example, 3D U-Net [4] extends convolutional and pooling operations into three dimensions, making it particularly effective for volumetric data such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) scans. Attention U-Net [5] integrates attention gates to automatically focus on the most relevant regions of interest, improving segmentation performance in cluttered or noisy images. Beyond CNN-based modifications, the U-Net framework has also been fused with other paradigms. For instance, transformer-based U-Nets [6] was introduced to better capture long-range dependencies and contextual information. Despite the success of U-Net and its derivatives [7], they fundamentally depend on the availability of large, expertly annotated training datasets.

### B. Strategies for Data-Scarce Scenarios

In order to mitigate the data scarcity, a common strategy is data augmentation, which synthetically increases dataset diversity by applying transformations such as rotation, scaling, flipping, elastic deformation, or intensity variations [8]. More advanced strategies include generative approaches to synthesize realistic training samples [9], [10]. However, synthetic images may fail to capture the full variability of real-world

pathologies and excessive or unrealistic transformations can introduce artifacts that reduce clinical plausibility [11]. Another widely adopted strategy is transfer learning, where a model is pre-trained on a large-scale dataset and subsequently fine-tuned on a smaller, domain-specific dataset [12]. However, this approach is often constrained by the domain gap between the source and the target medical datasets [13]. Also, fine-tuning still requires a sufficient number of examples to avoid overfitting, and its performance degrades sharply when only a handful of labeled samples are available.

To address data scarcity more directly, FSL has emerged as a variety of techniques enabling models to learn from a minimal number of labeled examples [2]. Within FSL, meta-learning (based on the concept of learn to learn) has become a dominant paradigm. Model-Agnostic Meta-Learning (MAML), proposed by Finn et al., is a prominent optimization-based meta-learning algorithm [14] that has achieved notable success in few-shot learning. However, its performance has been demonstrated primarily in image classification tasks, and most existing MAML-based frameworks remain centered on this domain [15]. In contrast, its potential for dense prediction problems such as medical image segmentation has not yet been fully explored [3]. Segmentation task presents additional challenges beyond classification, including substantial anatomical variability, severe class imbalance, and the requirement for fine-grained, pixel-level precision. This work aims to contribute to filling this gap by providing a direct empirical study of MAML on a multi-organ segmentation task.

### III. THE MAML APPROACH

As an optimization-based meta-learning approach, MAML's goal is to train a *meta-learner* (the outer-loop optimization algorithm) to produce a highly adaptable model, which we refer to as the *adapter*. The entire process is split into two phases: meta-training and meta-testing.

During meta-training, the meta-learner is exposed to a large number of distinct learning problems, drawn from a distribution of tasks  $p(\mathcal{T})$ . A task  $\mathcal{T}_i$  represents a specific learning problem, for example, "segment the liver" or "classify this skin lesion". For each task, a small dataset is provided, which is split into two disjoint sets: 1) a support Set ( $D_{\text{support}}$ ) of  $K$  labeled examples ( $K$ -shots) that the adapter uses to learn the specific task and 2) a query Set ( $D_{\text{query}}$ ) of samples used to evaluate how well the adapter has learned the task after being trained on the support set. The loss on this set provides the training signal for the meta-learner.

The training process is structured as a bi-level optimization problem, consisting of an inner loop for task-specific adaptation and an outer loop for meta-optimization across tasks. Let the adapter model be parameterized by a vector  $\theta$ .

*a) Inner Loop (Task-Level Adaptation):* For each task  $\mathcal{T}_i$  sampled from the meta-training set, a temporary, task-specific parameter vector  $\theta'_i$  is created. This is achieved by updating the initial parameters  $\theta$  for one or more gradient descent steps

using the loss calculated on that task's *support set*,  $D_{\text{support}}^{(i)}$ . For a single gradient step, this update is:

$$\theta'_i = \theta - \alpha \nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_{\theta}), \quad (1)$$

where  $\alpha$  is the inner learning rate,  $\mathcal{L}_{\mathcal{T}_i}$  is the task-specific loss function (e.g., cross-entropy) and  $f_{\theta}$  is the model (e.g. CNN) with  $\theta$  parameters.

*b) Outer Loop (Meta-Optimization):* The performance of the adapted parameters  $\theta'_i$  is then evaluated on the corresponding *query set*,  $D_{\text{query}}^{(i)}$ . The obtained loss serves as the meta-loss for that task. The initial parameter vector  $\theta$  is then updated by taking a gradient step with respect to this meta-loss, aggregated over a batch of tasks. The meta-update rule is:

$$\theta \leftarrow \theta - \beta \nabla_{\theta} \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta'_i}), \quad (2)$$

where  $\beta$  is the meta learning rate. Crucially, the gradient  $\nabla_{\theta}$  is calculated with respect to the initial parameters  $\theta$ , not the adapted parameters  $\theta'_i$ . This means that the optimization must differentiate through the inner-loop update (Eq. (1)). This "gradient through a gradient" operation requires computing second-order derivatives (Hessian-vector products), which allows the meta-learner to update the initial weights  $\theta$  by considering how the inner-loop update will affect the final query set loss. This process aims to find a shared initialization  $\theta$  that is positioned in a way that facilitates effective and efficient fine-tuning for a variety of novel tasks. In the meta-testing phase, a new task  $\mathcal{T}_{\text{test}}$  is sampled from outside the meta-training distribution. The model is initialized with the meta-learned parameters  $\theta$  and adapted to the new task using a small support set  $D_{\text{support}}^{(\text{test})}$ . Then, the performance is assessed on the corresponding query set  $D_{\text{query}}^{(\text{test})}$ , providing an evaluation of the model's ability to rapidly generalize.

### IV. EXPERIMENTAL STUDY ON ORGAN SEGMENTATION

#### A. Dataset description

All experiments were conducted on the BTCV Multi-Organ Abdominal CT Dataset [16]. It consists of 30 abdominal 3D CT volumes from healthy subjects. Each volume contains between 85 and 198 axial slices of  $512 \times 512$  pixels, with an anisotropic voxel spatial resolution of  $(0.54 \times 0.98 \times [2.5 \sim 5.0]) \text{ mm}^3$ . The dataset provides dense, manual annotations for 13 different abdominal organs, making it a suitable benchmark for creating a diverse task distribution. To evaluate the results on classes with various shapes, locations and textures we considered a subset containing left kidney, right kidney, spleen and liver, as suggested in [17].

#### B. Preprocessing pipeline

To prepare the data for our models, we followed a standardized preprocessing pipeline for each 3D volume:

- 1) Slice extraction: The 3D NIfTI volumes were processed into a collection of 2D axial slices. In order to mitigate the impact of class imbalance, we kept only slices in

which the organ mask occupied at least 10% of the image area for all the experiments.

- 2) Intensity clipping and normalization: The raw CT intensity values, measured in Hounsfield Units (HU), were clipped to a window of  $[-175, 250]$  HU. This step focuses the intensity range on soft tissues, which is relevant for abdominal organ segmentation. Following clipping, the intensity values were then normalized to a floating-point range of  $[0, 1]$ , as suggested in [18].
- 3) Resizing: all 2D slices and their corresponding masks were resized to a uniform dimension of  $256 \times 256$  pixels.

### C. Evaluation protocol

A strict data splitting protocol was designed to ensure a fair evaluation of the few-shot learning capabilities. This protocol involved splits at the patient, class (organ), and episodic levels. The dataset is randomly partitioned into a *meta-training pool* of 25 subjects and a held-out *meta-testing pool* of 3 subjects.

a) *Meta-Training*: For the meta-training phase, from the 25 patients, we used a group of 22 patients as the primary source for generating the *patient support set*, and the remaining 3 patients were designated for generating the *patient query set*. To simulate the few-shot learning scenario, we adopted a leave-one-out cross-validation scheme at the class level. In each experiment, 3 of these organs were used to form the task distribution in the meta-training and the left organ was reserved for meta-testing to evaluate the model's ability to adapt to a novel, unseen organ. The meta-training process operates by constructing *episodes*, where each episode simulates a few-shot learning problem for a specific task (e.g., "segment spleen"). To create an episode, a set of slices is sampled from the pool of all available data for that organ to create : 1) an *organ support set*, containing  $K$  examples (e.g.,  $K = 5$  for 5-shot configuration), drawn from the *patient support set* pool, which is used to adapt the model to the specific task during the inner loop; and 2) a larger, disjoint *organ query set*, drawn from the *patient query set* pool, which is used to evaluate the adapted model's performance and to compute the loss for the outer loop update.

b) *Meta-Testing*: For the meta-testing phase, the 3 dedicated held-out patients were used. This set was further split: one patient served as the source for the training data (support set), another for validation, and the last for testing (query set). This strict separation ensures a fair evaluation of generalization to both new patients and new organs.

### D. Model architecture

A standard 2D U-Net architecture was employed as the backbone for all experiments. It consists of a symmetric encoder-decoder structure. The encoder contains four down-sampling blocks, where each block applies two consecutive  $3 \times 3$  convolutions followed by a  $2 \times 2$  max-pooling operation. The decoder path mirrors the encoder, using transposed convolutions to up-sample the feature maps while halving the channel count at each level. This architecture was used for both the conventionally- and meta-trained models in order to

isolate the impact of the learning paradigm and ensures that any performance differences are attributable to the training strategy itself, not the model's capacity.

## V. RESULTS AND DISCUSSION

We present in this section the evaluation of our meta-learning framework. Since the dataset includes four organs, we consider four experimental configurations, each corresponding to a leave-one-organ-out setting, where a single organ is held out for meta-testing while the others are used for meta-training and meta-validation. We illustrate the training losses for each meta-model and its associated adapters across the four configurations in Figure 1. The reported results were obtained

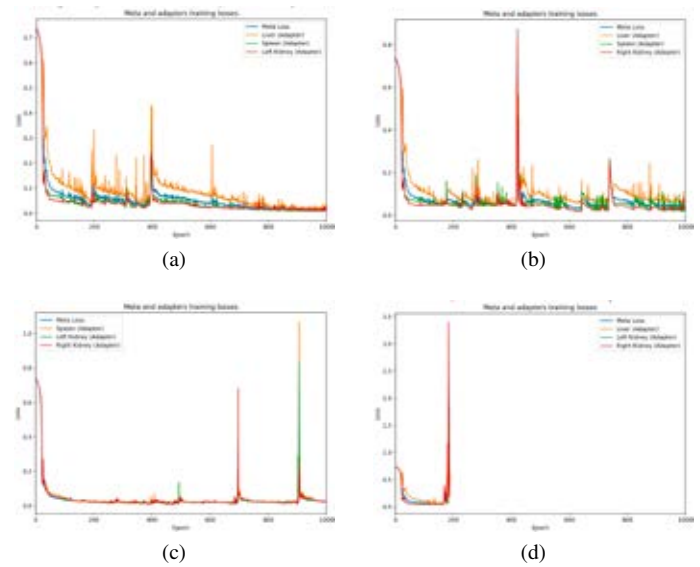


Fig. 1: Meta-training loss dynamics of the meta-model and adapters across four leave-one-organ-out configurations: (a) RK held-out (trained on L, S and LK), (b) LK held-out (trained on L, S and RK), (c) L held-out (trained on S, RK and LK), (d) S held-out (trained on L, RK and LK). Abbreviations: RK = Right Kidney, LK = Left Kidney, L = Liver, S = Spleen.

using the following hyperparameters: 1000 training epochs, 5 episodes per iteration, and 5 shots per class. Binary cross-entropy was employed as the training loss, while the Dice score served as the evaluation metric.

The training dynamics show that while the meta-learning process can converge stably for some organ combinations (notably when training on liver–spleen–left kidney), others (particularly liver–kidneys without spleen) are prone to instability, with the kidneys repeatedly emerging as more difficult organs for adaptation during meta-training. This behavior can be explained by organ-specific characteristics: kidneys are smaller, more variable in position and shape, providing weaker and noisier gradients during training, making them harder tasks for adaptation. In contrast, the spleen, being larger than kidneys and more homogeneous than liver, tends to stabilize optimization when included among the training tasks,

which may account for the smoother convergence observed in configurations where it is present.

For the meta-testing, we select the first configuration (liver–spleen–left kidney) at 900 epochs as it is the most stable. Figure 2 reports the fine-tuning performance of the meta-model, a randomly initialized model, and models pretrained on each of the training organs, evaluated with 3, 5, and 10 shots over 5–20 adaptation steps. Results show clear differences

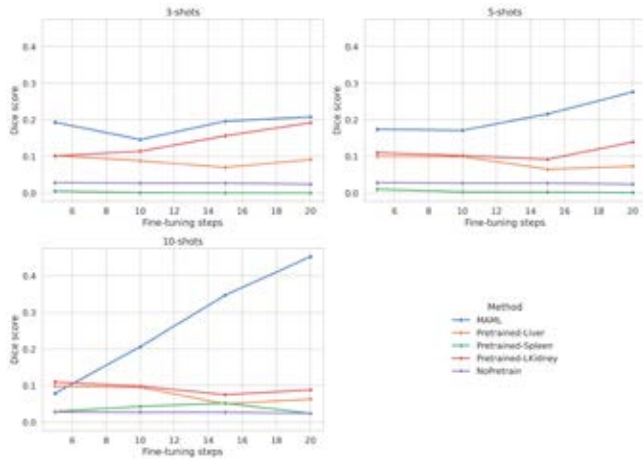


Fig. 2: A comparison of fine-tuning performance between meta-learned and conventional models.

across methods and shot settings. With only 3 shots, MAML reaches Dice  $\approx 0.19$ – $0.21$ , outperforming all classical transfers, except left-kidney pretraining ( $\approx 0.19$  at 20 steps). At 5 shots, MAML remains ahead (up to  $0.27$ ), while classical models stay below  $0.14$ . At 10 shots, MAML briefly lags at 5 steps ( $0.079$  vs.  $\sim 0.11$ ) but rapidly surpasses all baselines, reaching  $\sim 0.45$  by 20 steps whereas classical variants remain  $\leq 0.088$ . Overall, MAML consistently adapts faster and achieves higher performance, while classical pretraining offers modest gains.

## VI. CONCLUSION AND PERSPECTIVES

This study demonstrates that meta-learning, specifically MAML, can substantially improve the adaptation of segmentation models to new tasks (here, organs) in few-shot settings. Experiments on the BTCV dataset reveal that convergence stability is strongly influenced by the choice of training organs; nevertheless, under stable configurations MAML consistently outperforms a conventional U-Net baseline, achieving higher Dice scores across 3-, 5-, and 10-shot scenarios.

In future work, we will investigate the configurations where meta training exhibited unstable optimization and convergence difficulties. Understanding and mitigating these optimization challenges is crucial to ensure the robustness of the framework across all task configurations. We intend to consider more challenging datasets, including cases with tumors or lesions in different anatomical regions, and also datasets acquired under different institutions or patient populations. This is a critical step toward real-world deployment, as domain shifts caused by acquisition protocols or demographic variability often lead to

significant performance drops. We will also investigate cross-modality adaptation (such as CT to MRI), which is particularly relevant in clinical practice where data availability often varies across modalities. We also aim to extend our framework to 3D architectures for volumetric segmentation. This transition is non-trivial, as voxel-based representations substantially increase computational and memory requirements, making the optimization more challenging.

The code supporting this work is maintained in a private Git repository and can be provided upon reasonable request.

## REFERENCES

- [1] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” *LNCS*, vol. 9351, pp. 234–241, 10 2015.
- [2] E. Pachetti and S. Colantonio, “A systematic review of few-shot learning in medical imaging,” *Artificial Intelligence in Medicine*, vol. 156, p. 102949, 08 2024.
- [3] S. Luo, Y. Li, P. Gao, Y. Wang, and S. Serikawa, “Meta-seg: A survey of meta-learning for image segmentation,” *Pattern Recognition*, vol. 126, p. 108586, 02 2022.
- [4] Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, “3D U-Net: Learning dense volumetric segmentation from sparse annotation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2016.
- [5] A. Al Qurri and M. Almekkawy, “Improved U-Net with attention for medical image segmentation,” *Sensors*, vol. 23, no. 20, p. 8589, 2023.
- [6] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang *et al.*, “Transunet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers,” *Medical Image Analysis*, p. 103280, 2024.
- [7] X. Liu, L. Song, S. Liu, and Y. Zhang, “A review of deep-learning-based medical image segmentation methods,” *Sustainability*, vol. 13, p. 1224, 01 2021.
- [8] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of Big Data*, vol. 6, p. 60, 2019.
- [9] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, “Gan-based synthetic medical image augmentation for increased CNN performance in liver lesion classification,” in *Neurocomputing*, vol. 321. Elsevier, 2018, pp. 321–331.
- [10] J. Ehrhardt and M. Wilms, “Chapter 8 - autoencoders and variational autoencoders in medical image analysis,” in *Biomedical Image Synthesis and Simulation*, ser. The MICCAI Society book Series, N. Burgos and D. Svoboda, Eds. Academic Press, 2022, pp. 129–162.
- [11] L. Perez and J. Wang, “The effectiveness of data augmentation in image classification using deep learning,” in *arXiv preprint :1712.04621*, 2017.
- [12] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, “Convolutional neural networks for medical image analysis: Full training or fine tuning?” *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.
- [13] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, “Transfusion: Understanding transfer learning for medical imaging,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.
- [14] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” 03 2017.
- [15] S. M. R. Arnold, P. Mahajan, D. Datta, I. Bunner, and K. S. Zarkias, “learn2learn: A library for meta-learning research,” *CoRR*, vol. abs/2008.12284, 2020. [Online]. Available: <https://arxiv.org/abs/2008.12284>
- [16] harrigr, “Segmentation outside the cranial vault challenge,” 2015. [Online]. Available: <https://repo-prod.prod.sagebase.org/repo/v1/doi/locate?id=syn3193805type=ENTITY>
- [17] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Qiu, and D. Rueckert, “Self-supervision with superpixels: Training few-shot medical image segmentation without annotation,” in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 762–780.
- [18] L. Zhou, H. Liu, J. Bae, J. He, D. Samaras, and P. Prasanna, “Self pre-training with masked autoencoders for medical image classification and segmentation,” in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, 2023, pp. 1–6.