Pattern Recognition 153 (2024) 110471

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Meta-learners for few-shot weakly-supervised medical image segmentation

Hugo Oliveira [a,b,*], Pedro H.T. Gama [c], Isabelle Bloch [d], Roberto Marcondes Cesar Jr. [a]

[a] Institute of Mathematics and Statistics, Universidade de São Paulo, São Paulo, Brazil
[b] Computer Science Department, Universidade Federal de Viçosa, Viçosa, Brazil
[c] Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
[d] Sorbonne Université, CNRS, LIP6, Paris, France

ARTICLE INFO

Keywords:
Meta-learning
Weakly supervised segmentation
Few-shot learning
Medical images
Domain generalization

ABSTRACT

Most uses of Meta-Learning in visual recognition are very often applied to image classification, with a relative lack of work in other tasks such as segmentation and detection. We propose a new generic Meta-Learning framework for few-shot weakly supervised segmentation in medical imaging domains. The proposed approach includes a meta-training phase that uses a meta-dataset. It is deployed on an out-of-distribution few-shot target task, where a single highly generalizable model, trained via a selective supervised loss function, is used as a predictor. The model can be trained in several distinct ways, such as second-order optimization, metric learning, and late fusion. Some relevant improvements of existing methods that are part of the proposed approach are presented. We conduct a comparative analysis of meta-learners from distinct paradigms adapted to few-shot image segmentation in different sparsely annotated radiological tasks. The imaging modalities include 2D chest, mammographic, and dental X-rays, as well as 2D slices of volumetric tomography and resonance images. Our experiments consider in total 9 meta-learners, 4 backbones, and multiple target organ segmentation tasks. We explore small-data scenarios in radiology with varying weak annotation styles and densities. Our analysis shows that metric-based meta-learning approaches achieve better segmentation results in tasks with smaller domain shifts compared to the meta-training datasets, while some gradient- and fusion-based meta-learners are more generalizable to larger domain shifts. Guidelines learned from the comparative performance assessment of the analyzed methods are summarized to support those readers interested in the field.

1. Introduction

Despite the widespread use of deep learning in medical imaging, neural networks are still subject to a series of limitations that hamper their use in most real-world medical settings. The main hurdle in implementing Deep Neural Networks (DNNs) [1] into clinical practice is the data-driven nature of these models, which usually require hundreds, or even thousands of samples per class to fit properly, risking to suffer from underfitting in small-data scenarios. A compounding factor to this problem is the presence of domain shifts in real-world cases. In medical imaging, domain shift can be introduced to a task due to changes in imaging equipment or settings, differences in the cohort of training/test samples, imaging and label modalities, etc. Regarding segmentation tasks, an additional complication in using deep learning is the cost of producing the labels to be fed to the algorithms, as dense pixelwise labels are known to be expensive and require highly specialized anatomical knowledge from the physician. In this case, sparse labels fed to weakly supervised segmentation algorithms can be a good compromise between annotation cost and performance [2]. Still, the majority of works on one-/few-shot image segmentation do not consider the case of sparsely labeled images, only focusing on the case of fully labeled support sets [3,4].

The goal of the paper is to introduce a novel approach for Few-shot Weakly supervised Segmentation (FWS) of medical images in different modalities. In other words, we are interested in mitigating the limitations of semantic image segmentation with DNNs in problems with small-data, weak supervision for segmentation, and completely distinct image and label spaces during training and testing.

One common key point in few-shot learning is to introduce some form of prior knowledge into the models. A simple solution to this is to pretrain the model on a larger dataset. In RGB image domains, there are several datasets with thousands or even millions of annotated samples, which can be used to produce feature extractors to be leveraged in order to achieve good segmentation results with very few samples. Recently, Self-Supervised Learning (SSL) has been proven

* Corresponding author at: Computer Science Department, Universidade Federal de Viçosa, Viçosa, Brazil.
E-mail address: hugo.n.oliveira@ufv.br (H. Oliveira).
URL: https://sites.google.com/view/oliveirahugo (H. Oliveira).

https://doi.org/10.1016/j.patcog.2024.110471
Received 14 April 2023; Received in revised form 26 March 2024; Accepted 1 April 2024
Available online 8 April 2024

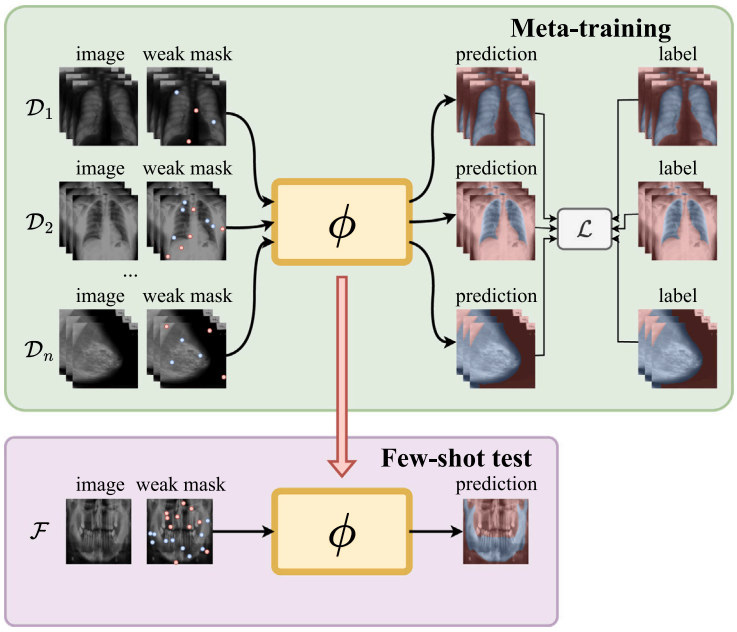


Fig. 1. Overview of a meta-learner ϕ for FWS being pretrained in an episodic fashion on multiple related tasks $\{D_1, D_2, \ldots, D_n\}$ pertaining to a meta-dataset (top) and being deployed to a target few-shot task $\mathcal{F}$ with a sparsely annotated support set.

to be a more robust initialization to such tasks [5] than supervised ImageNet pretraining. However, such strategies are often limited to natural images, not presenting considerable gains compared to random initializations for domains such as medical imaging. Another approach to introduce prior knowledge is to use meta-learning training. Already being successfully used in other pattern recognition tasks, it has been employed in few-shot image classification [6], with recent approaches for semantic segmentation gaining popularity [7]. However, the majority of works rely on ImageNet pretraining as feature extractors, while also not conducting tests on weak labels. In this scenario, there is currently a gap in reliable methods for FWS on non-RGB images.

We propose a framework capable of strong pretraining using Meta-Learning on radiological images that does not assume task-specific priors, does not require previous pretraining of the backbone prior to our meta-training phase, and, therefore, can be generalized to any FWS task in radiology. An overview of the proposed framework can be seen in Fig. 1. During its meta-training phase, it uses a meta-dataset $D = \{D_1, D_2, \ldots, D_n\}$, then it is deployed on an out-of-distribution (OOD) few-shot target task $\mathcal{F}$, where the single highly generalizable model ϕ, trained via a selective supervised loss function $\mathcal{L}$, is used as a predictor. As discussed in the following sections, ϕ can be trained in several distinct ways, such as second-order optimization, metric learning, and late fusion. Differently from Domain Adaptation (DA) tasks, the solutions described in this work concern tasks wherein the target domain is related to the meta-training data, but not available during training. Our tasks are closer to Domain Generalization [8] scenarios, since our framework uses pixelwise labeled data from multiple radiology domains (chest X-rays, mammographies, dental X-rays, etc.) aiming to be generalizable to other medical imaging tasks (e.g. computed tomography, magnetic resonance, positron emission tomography, etc.).

While the meta-learners used in this study are not novel by themselves, their application in Few-shot Weakly supervised Segmentation (FWS) is not straightforward. Thus, the main contributions of this work are the proposal of meta-learning algorithms to FWS and a thorough evaluation of the performance of multiple meta-learners in radiological image segmentation. We highlight the other contributions of this work as:

- A new generalizable FWS method that may incorporate meta-learners from multiple image classification paradigms;
- Relevant improvements of existing learners in real-world scenarios achieved by our proposal of SSL pre-training coupled with meta-learning;
- Guidelines and lessons learned from (1) detailed performance analysis of the multiple Meta-Learning algorithms being adapted to multiple FWS in 2D radiology; (2) analysis of multiple annotation styles and sparsity parameters.

This work is organized as follows. Section 2 reviews the literature and discusses the taxonomy adopted in this paper to compare the different approaches. Section 3 describes the assessed meta-learning methods while the experimental setup is explained in Section 4. The experimental results are presented and discussed in Section 5, with conclusions presented in Section 6.

2. Related work

2.1. Meta-learning for visual recognition

Meta-learning methods leverage multi-task learning in order to improve generalization for OOD tasks in image classification [6]. The idea is that, by generalizing for multiple datasets and tasks in the meta-training phase, the model is more well-equipped to deal with fully novel unseen tasks in the deployment phase. We highlight three distinct approaches for achieving Meta-Learning that are important to this paper, despite other paradigms (such as black-box or Bayesian approaches) also being common: (1) gradient-based – or optimization-based – methods, which acquire task-specific parameters via optimization of first- [9,10] or second-order derivatives [9,11,12]; (2) metric learning [13,14], wherein instead of directly predicting class probabilities, methods focus on learning distances across samples from similar and dissimilar classes; (3) fusion-based approaches, which leverage the intermediary representations of the support set to guide the predictions in the query set [15–17] via identity mapping (i.e. concatenation, multiplication, addition, etc.).

Gradient-based methods yield – often through second-order optimization – specific models for each task $\mathcal{T}$ in a meta-batch. Each subtask $\mathcal{T}$ computes its own temporary parameters, optimized using the task loss $\mathcal{L}_{\mathcal{T}}$. Finn et al. [9] introduce the MAML algorithm, a second-order framework, which updates its model parameters by taking

averages of cost's gradients of the specific task models evaluated on new samples of the task $\mathcal{T}$. Nichol et al. [10] propose the Reptile algorithm, which only uses first-order gradient information by updating its weights in the direction of the difference between task-specific parameters and global parameters. Also optimized via second-order derivatives, MetaSGD [11] aims to automatically learn step sizes for the SGD optimizer besides the model parameters. Raghu et al. [12] introduced another second-order algorithm similar to the MAML. Named ANIL, the method follows the MAML algorithm, with the novelty that in the *inner* loop, instead of updating all the parameters, this strategy only updates the ones related to the network output head, e.g. the last classification layers.

Metric-based approaches train a single model ϕ in multiple tasks $\mathcal{T}$. The objective of these approaches is to obtain an agnostic mapper to an embedding space where similar samples are closer than dissimilar ones. Snell et al. [13] propose the Prototypical Networks (ProtoNets), a model that tries to learn an embedding function that computes *prototypes* to a class (i.e. a d-dimensional vector that represents a class) and uses the distance to these prototypes for inference. For each class, features extracted from samples of their support – the labeled set of images of a task – are averaged to create its prototype vector. During inference, the feature extracted from a query image is compared with the prototypes, and the class of the closest prototype, according to some distance metric, is assigned to the query.

Fusion-based approaches, on the other hand, learn a single model ϕ where information of support sets is used to enhance the prediction of the query images. Similarly to metric-based methods, fusion-based approaches for meta-learning rely on an internal embedding from a neural network. However, instead of computing cross-sample similarities through a distance function (e.g. Euclidean, cosine, etc.) on the embeddings, such methods perform some form of late fusion (e.g. concatenation, addition, multiplication, cross-attention, etc.) on the support embeddings/labels and query samples. The Ridge Regression Differentiable Discriminator (R2D2) [16] uses the support embeddings and labels to train a fully tensorial logistic or ridge regressor using least-squares, which admits closed-form solutions. Similarly, MetaOptNet [15] leverages a highly discriminative embedding generated from a neural network to train a differentiable SVM. In both methods, the regressor obtained from the support data is then applied to the query samples through a simple matrix-matrix product, resulting in a few-shot classifier guided by the few labeled support examples.

2.2. Meta-learning for image segmentation

Recently, few-shot segmentation with meta-learning has evolved to alleviate the burden of the learners by filtering out irrelevant classes [18] and mitigate the bias of learners to base classes seen during meta-training [19], however most methods still do not fully explore weakly labeled support sets with a wide range of annotation modalities.

The most successful methods specifically designed for FWS are Guided Nets [17] and PANets [14]. Guided Nets rely on pretrained backbones to extract features from both support and query data, and apply late fusion in these embeddings to guide the prediction over the query set from the support codes. By contrast, PANets rely on a framework similar to ProtoNets [13] to compute prototypes for each class in the embedding space of the support set instead of leveraging late fusion to guide the prediction over the query. PANets also introduce Prototype Alignment Regularization (PAR) in the training phase for better label efficiency, wherein the query labels are also used to compute prototypes to predict the segmentations of support samples. Similarly to PANets, ProtoSeg [20] also use prototypes for conducting FWS, but instead of repurposing pretrained backbones, this approach is trained directly on related tasks from scratch in order to allow for inference over non-RGB images.

Hendryx et al. [21] adapt the Reptile [10] and First-Order-MAML (FOMAML) [9] to the problem of semantic segmentation. Their major contribution, however, is introducing the EfficientLab architecture, which is a convolutional network for semantic segmentation. Weakly supervised Segmentation Learning (WeaSeL) [22] applies the well-known MAML second-order optimization-based framework [9] to FWS in medical imaging by training it directly on tasks related to radiology, with the downside of being less efficient than first-order approaches. For a more thorough discussion on the few-shot meta-learning segmentation methods discussed previously and other few-shot and/or weakly-supervised segmentation approaches, we refer the reader to recent surveys [3,4,7,23].

Even though multiple works on meta-learners specifically designed for segmentation have appeared during the last few years, there is still a gap in such methods that can work with weakly annotated images. Aiming at encouraging a larger use of Meta-Learning for FWS tasks, in the remainder sections of this text we propose generalizable pipelines for porting meta-learners designed for image classification to weakly supervised segmentation tasks and test these approaches in real-world medical tasks.

3. Meta-learners for weakly supervised segmentation

We use most of the problem definitions from Gama et al. [22]. We consider a training dataset $\mathcal{D}$ as a set of pairs $(\mathbf{x}, \mathbf{y})$ of images $\mathbf{x} \in \mathbb{R}^{H\times W\times B}$ with dimensions $H \times W$ and B bands/channels, and semantic labels $\mathbf{y} \in \mathbb{R}^{H\times W}$. For each batch fed to an algorithm is partitioned into two sets, named the *support* set ($\mathcal{D}^{sup}$) and the *query* set ($\mathcal{D}^{qry}$), such that $\mathcal{D}^{sup} \cap \mathcal{D}^{qry} = \emptyset$. We define a segmentation task $\mathcal{T}$ as a tuple $\mathcal{T} = \{\mathcal{D}^{sup}, \mathcal{D}^{qry}, t\}$ (or, $\mathcal{T} = \{D, t\}$), where t is a target class or set of classes. In our setting, all segmentation tasks are binary during both meta-training and deployment, so t is a single class that can be referred to as *positive/foreground* in opposition to the *negative/background* class. $\mathcal{D}^{qry} = \{\mathbf{x}^{qry}, \mathbf{y}^{qry}\}$ is composed of a set of images ($\mathbf{x}^{qry}$) and associated densely labeled segmentation ground truth ($\mathbf{y}^{qry}$), while the support set $\mathcal{D}^{sup} = \{\mathbf{x}^{sup}, \mathbf{y}^{sup}\}$ contains another subset of images from the same dataset D as $\mathcal{D}^{qry}$, but paired with a weakly supervised mask $\mathbf{y}^{sup}$. We employ several distinct strategies detailed in the supplementary material of this manuscript to procedurally acquire the weak labels $\mathbf{y}^{sup}$ from the dense segmentation masks of the meta-training datasets.

A few-shot semantic segmentation task $\mathcal{F}$ is a specific type of segmentation task. The difference is that the samples of $\mathcal{D}^{sup}$ have their labels sparsely annotated, and the labels in $\mathcal{D}^{qry}$ are absent or unknown during training/tuning. Moreover, the number of samples $k = |\mathcal{D}^{sup}|$ is small, e.g., 20, 10 or even less.

The problem is then defined as follows. Given a few-shot task $\mathcal{F}$ and a set of segmentation tasks $\boldsymbol{\mathcal{T}} = \{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_n\}$, we want to segment the images from $\mathcal{D}_{\mathcal{F}}^{qry}$ using information from tasks in $\boldsymbol{\mathcal{T}}$ and information from $\mathcal{D}_{\mathcal{F}}^{sup}$. We also require that no pair of image/labels of $\mathcal{F}$ is present in $\boldsymbol{\mathcal{T}}$, in order to ensure that the only semantic information about $\mathcal{F}$ is in its support annotations.

In order to teach the model through supervised inputs during the meta-training phase, we employ supervised loss functions $\mathcal{L}_{sup}$. As not all pixels in an FWS task are labeled, we leverage the pixelwise Selective Cross-Entropy (SCE) loss function [20,22] in its binary form to conduct our supervised training on the labeled pixels in a ground truth $\mathbf{y}$ and prediction logits $\hat{\mathbf{y}}$:

$$\mathcal{L}_{sce}(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{N}\sum_{j=1}^{N} \mathbb{1}_j \left[\mathbf{y}_j \log \hat{\mathbf{y}}_j + (1-\mathbf{y}_j)\log(1-\hat{\mathbf{y}}_j)\right]. \tag{1}$$

In Eq. (1), N is the number of labeled pixels in $\mathbf{y}$, j is an index iterating over all pixels, $\hat{\mathbf{y}}_j$ is the probability predicted for each pixel j of being classified as pertaining from the positive class, and $\mathbb{1}_j \in \{0, 1\}$ is a flag indicating whether a pixel j has a valid annotation or not. $\mathcal{L}_{sce}$ is applicable to gradient-, metric- and fusion-based methods, with only the form of computing the logits $\hat{\mathbf{y}}_j$ varying across the different paradigms. We employ the SCE loss function in our experiments for all methods, as we observed early on that other supervised segmentation

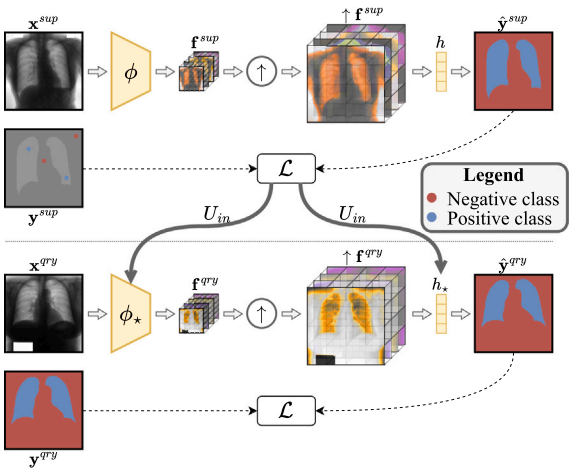


Fig. 2. Graphical illustration of one inner loop iteration for gradient-based FWS methods.

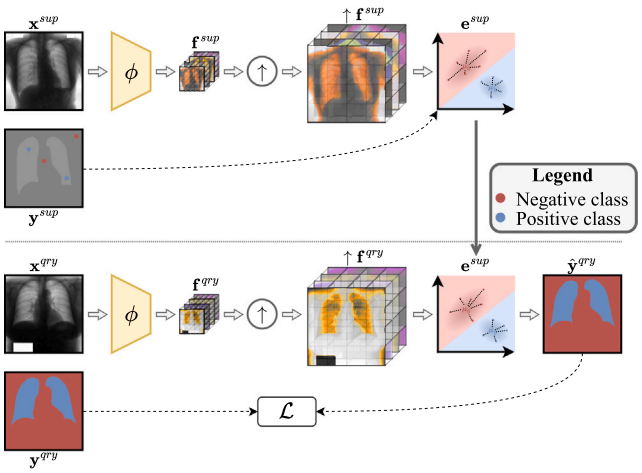


Fig. 3. Graphical illustration of one inner loop iteration for metric-based FWS methods.

loss functions (i.e. Dice [24], Focal [25], etc.) did not achieve the same performance as SCE.

Our proposed pipeline for FWS via Meta-Learning was conceived as a general strategy to convert algorithms originally designed for image classification to semantic segmentation, in theory being suitable to any gradient-, metric- or fusion-based method. We leverage knowledge gathered from previous works [20,22] on FWS using MAML [9] and ProtoNets [13] to generalize the frameworks to other state-of-the-art Meta-Learning algorithms [10–12,15,16]. Sections 3.1–3.3 discuss how each Meta-Learning paradigm can be ported to FWS tasks.

3.1. Gradient-based meta-learning for FWS

In order to adapt gradient-based meta-learners to FWS we employ FCNs and Encoder-Decoder architectures, and we divide each network architecture into two distinct parts: (1) a feature extraction component ϕ; and (2) a segmentation head h. ϕ receives support images $\mathbf{x}^{sup}$ and outputs embedded representations of the pixels in these images $\mathbf{f}^{sup} = \phi(\mathbf{x}^{sup})$, while h inputs $\mathbf{f}^{sup}$ and outputs segmentation predictions $\hat{\mathbf{y}}^{sup}$. For both FCNs and Encoder-Decoders, h is simply the last convolutional block responsible for the final pixelwise classification, while the feature extractor ϕ comprises all previous layers in the architecture – be it a sequence of Encoder/Decoder blocks or a CNN backbone.

The gradient-based pipeline can be observed graphically in Fig. 2 for a single inner loop. As one can see, a support image $\mathbf{x}^{sup}$ is initially fed through ϕ, generating features $\mathbf{f}^{sup}$, which are then upscaled using an interpolation function $\uparrow$ and fed to h, yielding a segmentation prediction $\hat{\mathbf{y}}^{sup}$. The prediction is then compared to the sparsely supervised ground truth $\mathbf{y}^{sup}$, from the support set, in the few labeled points available through $\mathcal{L}_{sup}$. The inner loop update function U_{in} operates on the gradients obtained through first- [10] or second-order [9,11] optimization, depending on the meta-learning algorithm of choice. U_{in} returns the task-specific feature extractor $\phi_\star$ and segmentation head $h_\star$, which are then fed with the query image $\mathbf{x}^{qry}$, resulting in a prediction $\hat{\mathbf{y}}^{qry}$ for the query set. $\hat{\mathbf{y}}^{qry}$ and $\mathbf{y}^{qry}$ are then compared through $\mathcal{L}_{sup}$, yielding gradients that can be backpropagated to meta-models ϕ and h.

In a real-world implementation of this idea, this procedure is repeated for all tasks randomly sampled in a meta-batch, each one with distinct $\phi_\star$ and $h_\star$ parameter sets, and yielding different gradients to be backpropagated to ϕ and h. The gradients can then be merged – usually via averaging – to update ϕ and h to more generalist parameter sets, highly adaptable to multiple tasks at once. This update to ϕ and h from the gradients obtained on the query set from the task-specific models is the outer loop of the optimization-based meta-learning algorithm. Additionally, it is highly desirable that the support image – or, realistically, support batch of images – is fed multiple times through ϕ and h, with the gradients being accumulated to generate $\phi_\star$ and $h_\star$ before feeding the query set to the task-specific models. The number of iterations through the support set used to compute $\phi_\star$ and $h_\star$ is a hyperparameter of gradient-based meta-learners, often being limited by the amount of memory in the GPU and/or training time constraints.

3.2. Metric-based meta-learning for FWS

As discussed in Section 2, metric-learning can also be used to achieve meta-learning through multi-task meta-training and some clever use of support set annotations [13,14,20] that are leveraged distinctly from an optimization-based approach. Instead of using the support set to tune task-specific models that should perform well on the query, metric-based methods instead use the labels from the support to compute prototypes in the embedding space, which can then be used as pivots to compute distances to query samples. This allows for extremely low-shot learning regimes (e.g. one-shot) and even zero-shot learning, the latter being unfeasible with most gradient-based methods.

We adapted metric-based meta-learners for FWS following the pipeline presented in Fig. 3 for a single inner loop iteration. Differently from most gradient-based approaches, the feature extractor ϕ remains frozen during the inner loops. The embedded spaces for the support $\uparrow \mathbf{f}^{sup}$ and query $\uparrow \mathbf{f}^{qry}$ sets, both in $\mathbb{R}^{C\times H\times W}$, are fundamental to pixelwise classification in this approach, assuming C output channels to ϕ. The few support labels available are used to compute pixel prototypes for each class, similarly to Snell et al. [13], which does this at an image level. As all tasks are binary in our approach, two centroids $\mu_0, \mu_1 \in \mathbb{R}^C$ are computed for the negative and positive classes, respectively; resulting in an embedded representation $\mathbf{e}^{sup}$. Prototypes only consider the labeled pixels from $\mathbf{y}^{sup}$, ignoring the unannotated ones from $\uparrow \mathbf{f}^{sup}$.

The upscaled query set features $\uparrow \mathbf{f}^{qry} = \phi(\mathbf{x}^{qry})$ are then projected onto the space $\mathbf{e}^{sup}$, where the distances of each query pixelwise feature vector can be computed in relation to μ_0 and μ_1 according to some distance metric d (e.g. Euclidean [13,20], cosine [14], Mahalanobis, Manhattan, etc.). Logits can then be computed for each query pixel according to their distances to the centroids of the negative and positive classes, allowing them to be fed to a supervised loss function $\mathcal{L}_{sup}$. The gradients obtained from $\mathcal{L}_{sup}(\mathbf{y}^{qry}, \hat{\mathbf{y}}^{qry})$ can then be backpropagated through the pipeline, reaching the trainable parameters ϕ. Similarly to the gradient-based methods, multiple inner loops such as the one shown in Fig. 3 are conducted in each meta-training iteration, with the gradients $\nabla_\phi \mathcal{L}_{sup}$ being added or averaged before updating ϕ on the outer loop.

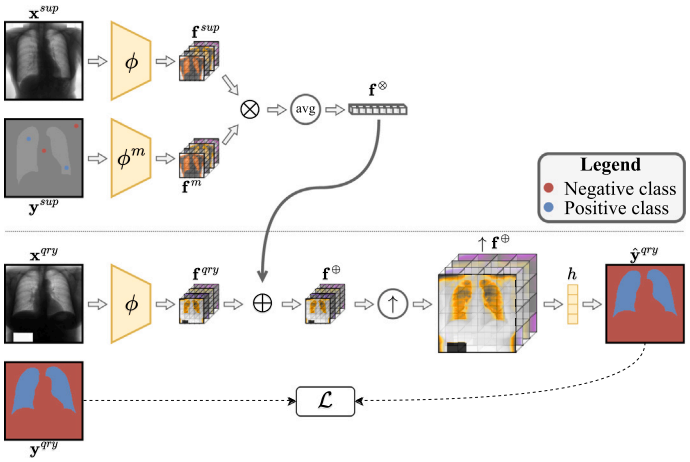


Fig. 4. Graphical illustration of one inner loop iteration for fusion-based FWS methods.

3.3. Fusion-based meta-learning for FWS

Fusion-based methods work by using the support set images and labels to "guide" [17,26] the classification or segmentation of the desired classes on the query set samples via late feature fusion between feature representations $\mathbf{f}^{sup}$ and $\mathbf{f}^{qry}$. Fig. 4 exemplifies the pipeline of a fusion-based algorithm along one inner loop, starting with the feature extraction from the support and query images via ϕ – yielding $\mathbf{f}^{sup}$ and $\mathbf{f}^{qry}$ – and the extraction of features from the sparse support mask $\mathbf{y}^{sup}$, resulting in $\mathbf{f}^{m}$. Feature representations $\mathbf{f}^{m}$ and $\mathbf{f}^{sup}$ are then fused using a function $\otimes$ and averaged into 1D vector $\mathbf{f}^{\otimes}$, responsible for guiding ϕ in segmenting query images according to the background and foreground classes in $\mathbf{y}^{sup}$. The guiding vector is then fused to the query features $\mathbf{f}^{qry}$ using another mapping $\oplus$ and resulting in $\mathbf{f}^{\oplus}$. The fused feature maps $\mathbf{f}^{\oplus}$ can then be passed through a segmentation head h, resulting in predictions $\hat{\mathbf{y}}^{qry}$, which is subsequently fed to $\mathcal{L}_{sup}$.

As shown in Fig. 4, some fusion-based approaches also extract features from the sparsely labeled support segmentation mask $\mathbf{y}^{sup}$ through a generic model referred to as ϕ^{m}. Guided Nets [17], for instance, repurpose the backbone ϕ to share parameters with ϕ^{m}, even though we found that keeping distinct parameters sets for ϕ and ϕ^{m} resulted in much more stable pretraining using this strategy. Another important aspect of fusion-based methods is the choice of the two functions: $\otimes$ – that merges support image and label features; and $\oplus$ – responsible for fusing the support and query feature representations. Common choices for these functions are concatenation, matrix multiplication, addition, pixelwise multiplication, or even trainable attention modules [3,4].

4. Experimental setup

Our experiments compare both well-known FWS algorithms in the literature (PANets [14], Guided Nets [17], WeaSeL [22] – referred to as MAML and ProtoSeg [20] – referred to as ProtoNets) and novel algorithms based on meta-learners designed for few-shot classification and ported to FWS (MetaSGD [11], ANIL [12], Reptile [10], R2D2 [16] and MetaOptNet [15]).

We explored different neural network architectures as backbones for the meta-learners, including: U-Net (**U**) [20,27], EfficientLab-6-3 (**E**) [21], DeepLabv3 (**D**) [28] and an FCN with a ResNet-12 (**R**) [15] backbone.

Aiming to improve the performance of meta-learners by introducing inductive bias learned directly from related data, we also tested CNNs pretrained with SSL previously to the meta-training phase. This strategy aims at leveraging pretraining without human annotations to improve the initial embeddings $\mathbf{f}^{sup}$ and $\mathbf{f}^{qry}$ for the support and query sets generated by ϕ, potentially improving the overall performance on the target FWS task during deploy. For that, we employed the SimSiam [29] algorithm for pretraining a ResNet-18 and a ResNet-50 [30] on the radiological datasets used during meta-training, resulting in a backbone ϕ^{SSL} and a pretext task prediction head h^{SSL}. We then employ ϕ^{SSL} as the starting point for our feature extractors ϕ and resume the meta-training phase.

As the performance of few-shot algorithms is notoriously variable according to the chosen support set samples, all results shown in Section 5 were computed according to a 5-fold cross-validation procedure, with paired samples for each fold in all algorithms. As all of our tasks are binary by experimental design, the simple Intersection over Union (IoU, also known as the Jaccard index) between the positive and negative classes was our main measure for assessing the performance of FWS meta-learners.

Our implementation of meta-learners for FWS was coded using the Pytorch[1] and learn2learn[2] libraries. Experiments were conducted in machines running RTX 2070 GPUs with 8 GB of memory, so many hyperparameters for the meta-learning algorithms (e.g. meta batch size, number of inner loops, number of filters in the architectures, adaptation steps in 2nd order gradient-based methods, etc.) were selected aiming to fit this capacity. Whenever possible, we set the default optimizer for our algorithms as Adam [31], unless in methods that use quadratic program solvers, such as MetaOptNet [15] or R2D2 [16], which rely on standard Stochastic Gradient Descent (SGD). As time limitation was our main consideration in designing the experiments shown in this work, we leveraged knowledge from previous works [20,22] to set most of the other hyperparameters that do not directly affect the memory usage of meta-learners. Readers can refer to our official implementation[3] for details.

4.1. Experiment organization

In order to test the efficacy of the meta-learners shown in Section 3 for FWS, we designed an experimental setup aiming to evaluate the performance of each algorithm in multiple radiological image segmentation tasks. We use the *points* weak annotation style to assess the performance of algorithms in very small data scenarios for two distinct few-shot tasks: $\mathcal{T}_{id}$ – the OpenIST dataset[4] in lung segmentation task (*OpenIST-lungs*); and $\mathcal{T}_{ood}$ – the Panoramic Dental X-rays [32] in inferior mandible segmentation task (*Panoramic-mandible*). Information on the source Chest X-ray (CXR), Dental X-ray (DXR), Mammographic X-ray

[1] https://pytorch.org/.
[2] http://learn2learn.net/.
[3] https://github.com/hugo-oliveira/fsws_metalearning.
[4] https://github.com/pi-null-mezon/OpenIST.

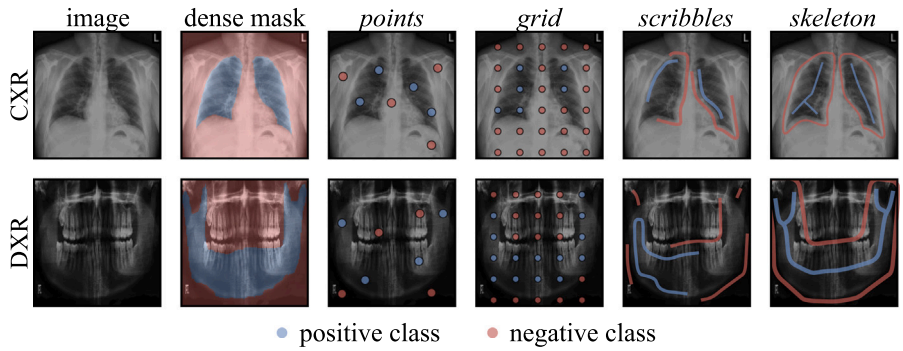


Fig. 5. Weak annotation styles (4 last columns) studied in this work for two distinct radiological tasks: (1) *OpenIST-lungs* (top row); and (2) *Panoramic-mandible* (bottom row).

(MXR), and Digitally Reconstructed Radiograph (DRR) datasets used in this research can be seen in the supplementary material accompanying this manuscript.

Task $\mathcal{T}_{id}$ represents a relatively "in-distribution" (ID) task very close to many domains used during meta-training, as both the Shenzhen/Mongomery sets and the Chest X-ray 14 dataset are very similar to the OpenIST data and also contain lung annotations. Task $\mathcal{T}_{ood}$, however, is a very out-of-distribution task compared to the meta-training domains, as no other dataset contains segmentation for the inferior mandible and only one other DXR dataset (IVisionLab) is used during the meta-training for the quite distinct task of teeth segmentation. To avoid the meta-learners simply overfitting on the target dataset [33], we hold out each target domain in an experiment from the meta-training phase, assuring that the first time the methods/networks have seen each target dataset is during the test phase.

Additionally to the experiments on ID and OOD 2D images, we tested our meta-learners pretrained on the 2D CXRs, MXRs, and DXRs in four tasks from three volumetric datasets: (1) StructSeg [34] on the *Both Lungs* and *Heart* tasks (*StructSeg-lungs* and *StructSeg-heart*); (2) Medical Imaging Segmentation Decathlon (MSD) [35] slices on the *Spleen* task (*MSD-spleen*); and (3) private data from Oliveira et al. [36] in *Cerebellum* segmentation on pediatric MRI data (*STAP-cerebellum*). We selected these tasks as targets due to the relatively large size of these organs in the images in comparison to the whole volumes. This selection was conducted in order to avoid the inherent difficulties of compensating for domain shift in datasets with imbalanced target classes [37].

While the meta-training is conducted with the largest variability possible for the choice of support/query set samples and weak support masks, we fixed the seeds of all randomly chosen variables in the sparsification algorithms for the support set of the target task $\mathcal{T}$, forcing them to be the same on a pixel-level for all samples.

To evaluate the performance of the best meta-learners in other weakly supervised segmentation styles, we conduct a series of experiments with *grid*, *scribbles*, and *skeleton* annotations, as depicted in Fig. 5. These experiments are presented in Section 5.2 for *Panoramic-mandible*, *OpenIST-lungs*, *StructSeg-lungs*, *StructSeg-heart*, *MSD-spleen* and *STAP-cerebellum*. At last, in Section 5.4 we exemplify how SSL pretraining [29] can benefit the OOD task performance of our meta-learners in the tasks of *MSD-spleen* and *StructSeg-lungs* for *points* and *scribbles* annotations.

5. Results and discussion

5.1. Results intra meta-learning paradigms

We present initially the results in the *points* annotation style for 1- and 5-shot scenarios, each one with 4 distinct label configurations: (1) 1 annotated point (p = 1) with a dilation radius of 1 (r = 1); (2) p = 1/r = 3; (3) p = 5/r = 1; and (4) p = 5/r = 3. These experiments correspond to the *Panoramic-mandible* task, where the domain shift is larger than *OpenIST-lungs*, as the focus of this work is in Domain Generalization instead of simple cross-dataset transfer learning. This results in 8 distinct annotation settings, which is the number of result columns in Tables 1–3. Bold values in result tables represent the best overall results for each label configuration (column), while underlined values highlight the backbones, which had the best performance for a meta-learner. All metrics within 0.01 of IoU difference from the best result are highlighted in those tables in order to point not only to the best overall algorithm/backbone pairs but also to other strategies with comparable performance. Additional results for meta-learning algorithms in *Panoramic-mandible* and *OpenIST-lungs*, including statistical confidence intervals, can be seen in the supplementary material that accompanies this manuscript.

We present results for algorithms in each Meta-Learning paradigm separately in order to sort the more label-efficient algorithms for comparisons cross-paradigms. Tables 1–3 show, respectively, the IoU metrics for the gradient-, metric- and fusion-based methods for the 4 backbones analyzed in this work, as well as the baseline with ResNet-12.

Table 1 shows IoU results for pure 2nd order [9,11], pure 1st order [10], as well as optimization-based strategies that use both 2nd and 1st order gradients [12]. ANIL-**D** achieved the best overall results in gradient-based methods, yielding the best performances in all FWS configurations. Hence, the hybrid strategies of mixing 1st order gradients to train the backbone and 2nd order gradients to train the segmentation head were the most label-efficient gradient-based methods by quite some margin. Most of this strategy's success can be attributed to the larger amount of adaptation steps that can be computed to the segmentation head (10 adaptation steps in ANIL) in comparison to applying 2nd order gradients to the whole backbone (2 adaptation steps for MAML and MetaSGD). Thus, with additional GPU memory, one could theoretically improve the performance of full 2nd order approaches in theory. However, we conducted early tests with up to 10 adaptation steps on MAML and MetaSGD and these approaches tended to apparently overfit on the few annotated pixels with more adaptation steps than 5, possibly due to the larger parameter capacity compared to the amount of labeled data being fitted by the 2nd order optimization in the whole backbone.

Metric-based methods shown in Table 2, in comparison to the baseline with ResNet-12 backbone, highlight the superiority of PANets [14] compared to ProtoNets [13] in all scenarios. The better performance of PANets can be attributed to two factors: the use of the cosine distance instead of Euclidean distance and the additional PAR regularization proposed by Wang et al. [14] – further explained in Section 2.2. EfficientLab outperformed other backbones in all but the two most sparsely labeled scenarios (1-shot/p = 1), while the other architectures performed quite well in 1-shot/p = 1/r = 1 and 1-shot/p = 1/r = 3, but lacked the capacity of learning from more annotations as well as the PANet-**E**.

Table 1
IoU results for four distinct gradient-based methods (MAML [9,22], MetaSGD [11], ANIL [12] and Reptile [10]) with the four segmentation backbones pretrained on our meta-dataset and tuned on the few-shot weakly-supervised support samples with the *points* annotation style.

Method		1-shot				5-shot			
		p = 1 r = 1	p = 1 r = 3	p = 5 r = 1	p = 5 r = 3	p = 1 r = 1	p = 1 r = 3	p = 5 r = 1	p = 5 r = 3
Baseline	**R**	.312	.311	.478	.482	.443	.450	.556	.558
MAML	**U**	.205	.239	.321	.360	.319	.336	.476	.473
	R	.260	.269	.387	.394	.373	.358	.249	.243
	E	.179	.185	.253	.254	.206	.207	.301	.304
	D	.328	.335	.424	.431	.322	.337	.488	.501
MetaSGD	**U**	.306	.351	.306	.304	.301	.300	.411	.435
	R	.268	.277	.269	.260	.340	.390	.441	.473
	E	.244	.246	.376	.388	.324	.320	.368	.366
	D	.222	.234	.173	.191	.256	.272	.421	.431
ANIL	**U**	.327	.323	.369	.385	.361	.366	.475	.473
	R	.187	.203	.216	.260	.314	.341	.218	.190
	E	.361	.364	.504	.506	.484	.490	.507	.517
	D	**.454**	**.451**	**.532**	**.541**	**.546**	**.552**	**.619**	**.627**
Reptile	**U**	.334	.341	.328	.326	.349	.355	.365	.386
	R	.203	.205	.336	.354	.394	.414	.524	.536
	E	.206	.210	.281	.318	.367	.387	.314	.305
	D	.160	.161	.210	.205	.085	.080	.283	.317

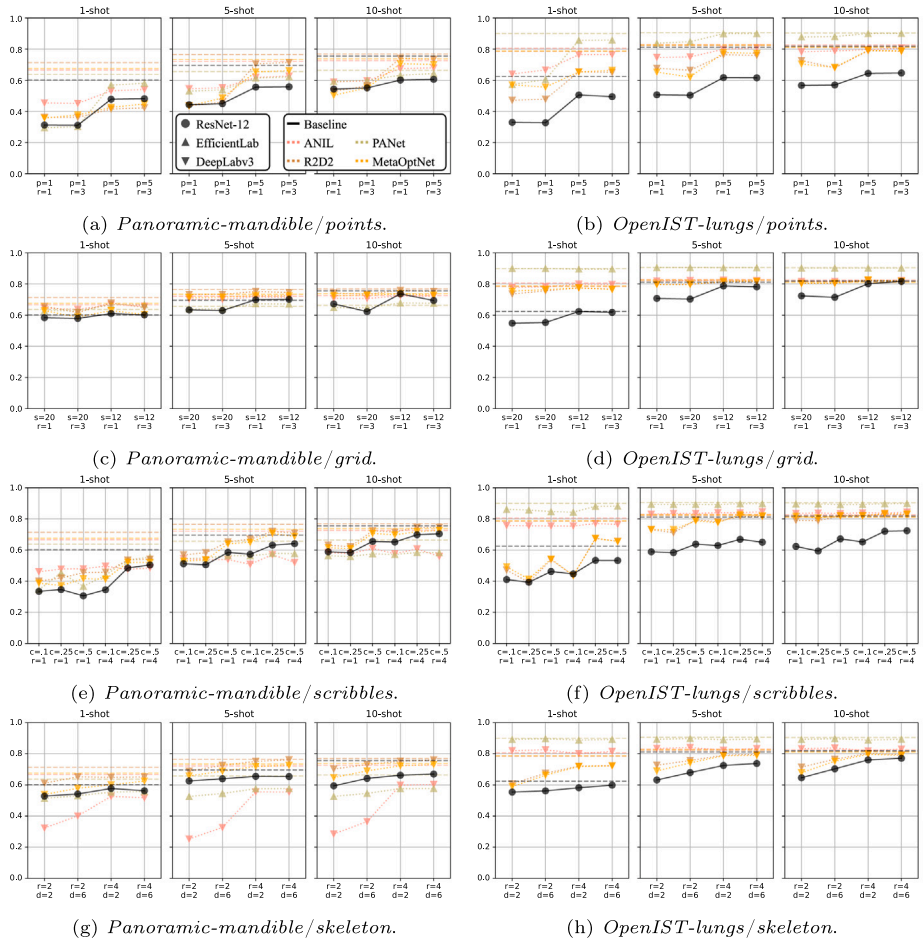
Table 2
IoU results for two distinct metric-based methods (ProtoNets [13,20] and PANets [14]) with the four segmentation backbones pretrained on our meta-dataset and tuned on the few-shot weakly supervised support samples with the *points* annotation style.

Method		1-shot				5-shot			
		p = 1 r = 1	p = 1 r = 3	p = 5 r = 1	p = 5 r = 3	p = 1 r = 1	p = 1 r = 3	p = 5 r = 1	p = 5 r = 3
Baseline	**R**	.312	.311	.478	.482	.443	.450	.556	.558
ProtoNet	**U**	.371	.365	.364	.362	.454	.462	.558	.563
	R	.363	.364	.397	.395	.460	.463	.533	.540
	E	.319	.336	.445	.453	.448	.449	.521	.525
	D	.382	.381	.403	.399	.458	.456	.556	.568
PANet	**U**	**.416**	**.411**	.419	.432	.511	.509	.554	.561
	R	**.423**	**.418**	.512	.510	.518	.522	.592	.600
	E	.295	.303	**.570**	**.581**	**.530**	**.539**	**.615**	**.622**
	D	**.419**	**.421**	.462	.462	.481	.484	.578	.584

Table 3
IoU results for three distinct fusion-based methods (Guided Nets [17], R2D2 [16], and MetaOptNet [15]) with the four segmentation backbones pretrained on our meta-dataset and tuned on the few-shot weakly supervised support samples with the *points* annotation style.

Method and backbone		1-shot				5-shot			
		p = 1 r = 1	p = 1 r = 3	p = 5 r = 1	p = 5 r = 3	p = 1 r = 1	p = 1 r = 3	p = 5 r = 1	p = 5 r = 3
Baseline	**R**	.312	.311	**.478**	**.482**	.443	.450	.556	.558
Guided Net	**U**	.015	.013	.020	.012	.000	.000	.000	.000
	R	**.450**	**.449**	.447	.443	.450	.450	.449	.447
	E	.295	.298	.286	.306	.272	.272	.272	.291
	D	.150	.161	.140	.164	.254	.259	.260	.265
R2D2	**U**	.320	.324	.292	.341	.390	.419	.531	.550
	R	.365	.365	.425	.399	.487	.502	.596	.614
	E	.268	.272	.385	.422	**.529**	**.526**	.677	.681
	D	.362	.362	.415	.423	.436	.460	**.706**	**.715**
MetaOptNet	**U**	.383	.379	.415	.408	.424	.423	.458	.458
	R	.390	.382	.400	.382	.494	.489	.627	.640
	E	.257	.262	.399	.444	.513	.515	.659	.670
	D	.358	.377	.426	.448	.435	.486	.654	.662

Guided Nets, mainly with a ResNet-12 backbone, had a strong start in extremely low-data scenarios (e.g. 1-shot/p = 1), but were unable to evolve to learn from more annotations, maintaining their performance close to 0.45 of IoU throughout all other experiments. As for R2D2 and MetaOptNet, their most promising backbones reach relatively similar performances in very sparsely annotated scenarios – from 0.36 to 0.39 in 1-shot/p = 1 – with the considerable advantage that these architectures can continue to learn from the larger amount of annotated data in 1-shot/p = 5 and 5-shot. We highlight that R2D2-**E**, R2D2-**D** and MetaOptNet-**E**, MetaOptNet-**D** show good performances whenever larger amounts of annotated support pixels are provided.



(a) *Panoramic-mandible/points.* (b) *OpenIST-lungs/points.*

(c) *Panoramic-mandible/grid.* (d) *OpenIST-lungs/grid.*

(e) *Panoramic-mandible/scribbles.* (f) *OpenIST-lungs/scribbles.*

(g) *Panoramic-mandible/skeleton.* (h) *OpenIST-lungs/skeleton.*

Fig. 6. IoU results for multiple weakly-supervised annotation styles in 1-, 5- and 10-shots for the best baseline backbone (**R**) and the two overall best algorithm/backbone pairs in each paradigm (ANIL-**D**, PANets-**E**, R2D2-**D** and MetaOptNet-**D**). Rows reflect distinct weakly-supervised annotation styles, respectively: *points*, *grid*, *scribbles*, *skeleton*. The left column represents results for the *Panoramic-mandible* task (large domain shift), while the rightmost column depicts results for the *OpenIST-lungs*. Dotted lines reflect the performance of segmentation algorithms on the weakly supervised support sets, while the dashed horizontal lines show the results tuned on the densely annotated support masks. Better viewed in color.

From the results presented in Tables 1–3, we chose 7 distinct algorithm/backbone pairs for further analysis in the following sections: Baseline-**R**, ANIL-**E**/**D**, PANets-**R**/**E**, R2D2-**D** and MetaOptNet-**D**.

5.2. Weak annotation styles

While Section 5.1 focused on results for *points* annotations, the present section shows results for three additional weakly supervised mask styles: *grid*, *scribbles* and *skeleton*.

Fig. 6 depicts some broader trends applied to all experiments. As expected, the *OpenIST-lungs* segmentation task – wherein the domain shift is smaller both in the pixel- and label-space compared to the meta-dataset – achieves considerably better segmentation performances than the *Panoramic-mandible* segmentation. Even in very sparsely labeled support sets (i.e. *points* in 1-shot/p = 1), some algorithms reach 0.6 or more of IoU, while the 5-shot/p = 1 scenario presents performances for PANets with IoU larger than 0.8. All other annotation styles *grid*, *scribbles* and *skeleton* in their most sparsely annotated scenarios even reach 0.9 of IoU for *OpenIST-lungs*. Similarly to results reported by Gama et al. [20], metric-based methods seem to be a much more suitable predictor for tasks in target domains where the domain shift from the meta-dataset is small. ANIL-**D** follows PANets quite closely in this task, while fusion-based approaches reached a lower ceiling in performance on *OpenIST-lungs*.

One should notice that for OpenIST only annotations in randomly selected *points* did not reach around 0.8 of IoU, while all of *grid*, *scribbles*, and *skeleton* did. This quickly reaching a ceiling in performance in all but one annotation style implies that *points* are a less label-efficient way of providing weak annotations.

Depending on the meta-learner, *points* in its more sparse setting (1-shot/p = 1) achieves between around 0.45/0.70 of IoU for the *Panoramic-mandible* and *OpenIST-lungs* tasks, respectively, while the highly efficient *skeleton* annotation style in its more sparse setting (1-shot/r = 2) yields 0.60/0.90 for MetaOptNet/PANets. In fact, annotations in *skeleton* and *scribbles* seem to be more efficient than *points*, mainly because the time spent in labeling single random points in an image is similar to the time spent in delineating a scribble or a skeleton for a commonly shaped organ, while also providing more annotated points to tune the learner into the target task. Both draw-oriented annotations also provide a relatively clear guide for the algorithm indicating the organ borders, while randomly sampled points do not provide this information.

Another clearly seen trend in Fig. 6 is the tiny performance gap between sparse (dotted lines) and dense (dashed horizontal lines) annotations. While the gap is considerably large in *Panoramic-mandible* for more sparsely annotated scenarios in all annotation styles, for *OpenIST-lungs* in all scenarios and *Panoramic-mandible* in more densely annotated scenarios it is negligible for the best meta-learners.

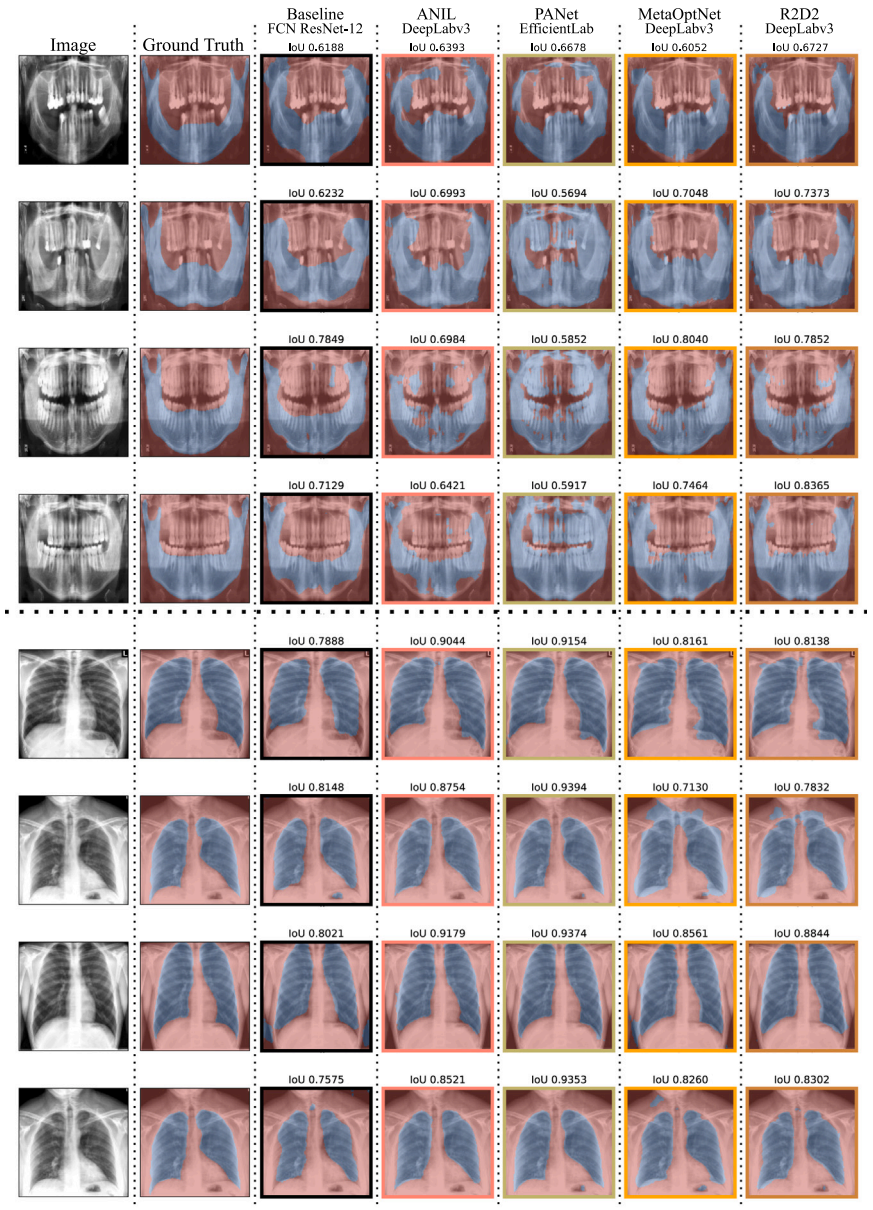


Fig. 7. Qualitative results for FWS on two target 10-shot tasks: *Panoramic-mandible* (top 4 rows) and *OpenIST-lungs* (bottom 4 rows); for 4 distinct annotation styles: *points* (1st and 5th rows), *grid* (2nd and 6th rows), *scribbles* (3rd and 7th rows) and *skeleton* (4th and 8th rows). The positive class is represented in blue and the negative class is shown in red. For each task/annotation style, we show one sample of the query set ($\mathcal{T}^{qry}$). Segmentation predictions and sample IoU metrics for 4 of the best meta-learners (ANIL-**D**, PANet-**E**, MetaOptNet-**D** and R2D2-**D**) and Baseline-**R** are shown. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 7 shows sample segmentation predictions from 4 meta-learners and a baseline in 2D radiology data. Following the trends shown in Fig. 6, qualitative segmentation predictions in the *Panoramic-mandible* task are better performed by the fusion-based methods (R2D2-**D** and MetaOptNet-**D**). As for the task with a small domain shift, the metric-based PANets-**E** work better in *OpenIST-lungs*, followed by the gradient-based ANIL-**D**. One can also derive from the 4 top lines in Fig. 7 that the *points* style is the least effective weak annotation modality for both *Panoramic-mandible* and *OpenIST-lungs*, while *scribbles* and *skeleton* labels show overall superior performance for the best meta-learners in each scenario.

5.3. Results for 2D slices from volumetric data

This section presents our experiments on 2D slices from 3D data, representing fully OOD tasks in relation to the meta-training datasets. Quantitative results for four tasks (*StructSeg-lungs*, *StructSeg-heart*, *MSD-spleen* and *STAP-cerebellum*) are shown in Fig. 8 for two annotation styles: *grid* and *skeleton*. As all 2D slice tasks are considerably more OOD compared to the meta-training datasets, we observe similar results to the ones presented for the *Panoramic-mandible* task in Fig. 6, with fusion-based methods achieving higher performance than their counterparts. These best results obtained by R2D2 and MetaOptNet are followed by the optimization-based ANIL, with the similarity-based PANets showing performance close to the baseline in most tasks and annotation styles. Weak *grid* annotations seemed to perform better than the skeleton labeling style in 2D slice tasks, possibly due to the spatial uniformity of such annotations in tasks with large domain shifts, even though it is often more time-demanding to label a grid of points rather than simply draw a skeleton and outline of an organ.

Another interesting trend shown in Figs. 8(b), 8(d), 8(f) and 8(h) is that the baseline method was clearly outperformed by all meta-learners

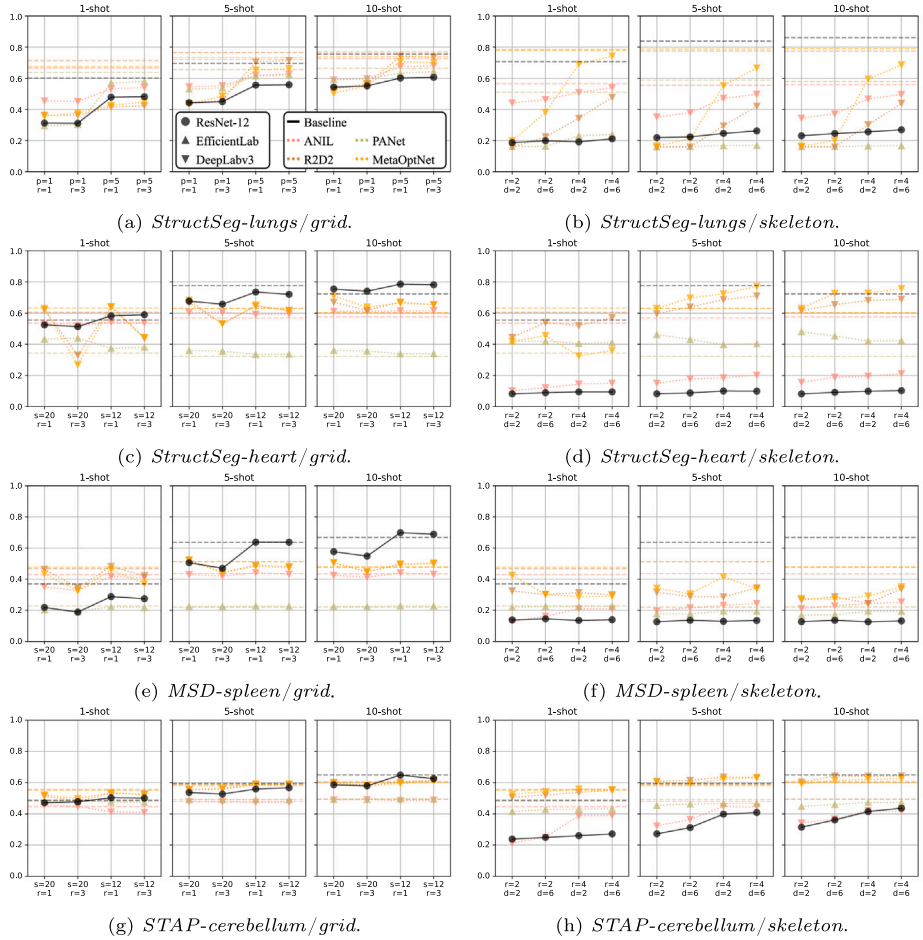


Fig. 8. IoU results for multiple weakly-supervised annotation styles in 1-, 5- and 10-shots for the best baseline backbone (**R**) and the two overall best algorithm/backbone pairs in each paradigm (ANIL-**D**, PANets-**E**, R2D2-**D** and MetaOptNet-**D**). Rows reflect distinct tasks, respectively: *StructSeg-lungs*, *StructSeg-heart*, *MSD-spleen* and *STAP-cerebellum*. The leftmost column represents results for *grid* annotations, while the right column depicts *skeleton* results. Dotted lines reflect the performance of segmentation algorithms on the weakly supervised support sets, while the dashed horizontal lines depict the results tuned on the densely annotated support masks. Better viewed in color.

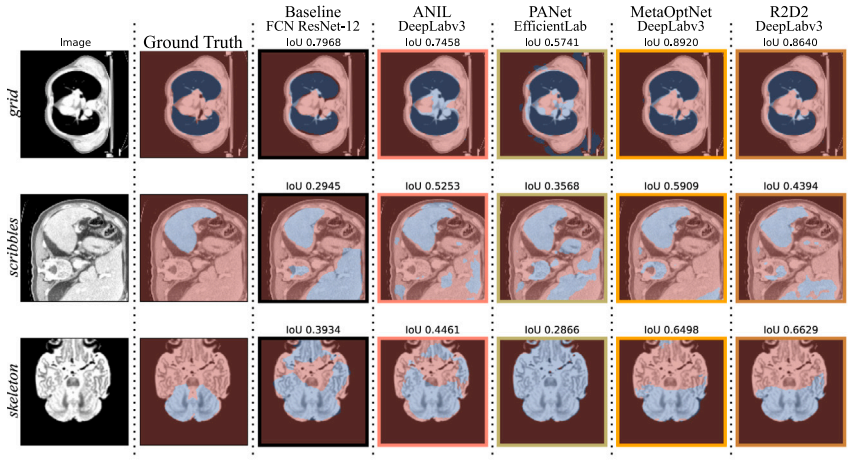


Fig. 9. Qualitative results for FWS on three target tasks: *StructSeg-lungs* on *grid* annotations (top row), *MSD-spleen* on *scribbles* annotations (middle row) and *STAP-cerebellum* on *skeleton* annotations (bottom row). Meta-learners and annotation styles are analogous to Fig. 7. Better viewed in color.

in the *skeleton* annotation style for all tasks, but PANets in the *StructSeg-lungs* task. In contrast to that, the baseline had a very competitive performance for *grid* annotations, even outperforming all meta-learners in most sparsity configurations in Figs. 8(c) (*StructSeg-heart*) and 8(e) (*MSD-spleen*). Again, we attribute this better performance of the baseline to the higher density of annotations in *grid*. This implies that the baseline is less adaptable than meta-learners in highly sparse and unorthodox annotation styles.

Table 4
IoU results for improved meta-learners in the *MSD-spleen* dataset for *points* (*num_points = 5, radius = 3*) and *scribbles* (*proportion = 0.25, thickness = 4*). Bold results represent the best results for each algorithm across network architectures and underscored values indicate the best overall results in a whole column.

Method		*Points*			*Scribbles*		
		Shots			Shots		
		1	5	10	1	5	10
Baseline	**R18+SS**	.145	.156	.219	.140	.166	**.173**
	R50+SS	**.176**	**.259**	**.318**	**.150**	**.180**	.154
ANIL	**DLabv3**	.337	.365	.385	.197	.233	.243
	R18+SS	.300	**.509**	**.477**	.235	.486	.459
	R50+SS	**.403**	.428	.375	**.444**	**.556**	**.470**
PANet	**EffLab**	.200	.211	.212	.190	.164	.170
	R18+SS	**.544**	**.348**	**.398**	**.215**	**.207**	**.213**
	R50+SS	.249	.231	.252	.189	.194	.189
R2D2	**DLabv3**	.391	.447	**.487**	**.285**	.370	.366
	R18+SS	.389	.457	.480	.201	.237	.252
	R50+SS	**.412**	**.470**	.477	.234	**.463**	**.411**
MetaOptNet	**DLabv3**	.342	.443	.463	.216	.315	.347
	R18+SS	.298	.470	.469	.183	.310	.334
	R50+SS	**.366**	**.492**	**.517**	**.272**	**.449**	**.456**

Finally, we show segmentation results for three of the 2D slice tasks in Fig. 9. As the domain shift is larger for these tasks, similar to the *Panoramic-mandible* task, R2D2 and MetaOptNet achieved the overall best FWS results in comparison to the baseline and gradient-/metric-based strategies. More specifically, the metric-based PANet performed particularly poorly in these high domain-shifted tasks, segmenting other organs as the structure of interest in the majority of cases. This implies that metric-based methods are unable to correctly deal with the inherent ambiguity of FWS tasks, possibly due to the difficulties in creating an embedding space that is discriminative to a wide range of real-world unseen OOD tasks.

5.4. Improving meta-learners with SSL

Additionally to the experiments with multiple algorithms within meta-learning paradigms (Section 5.1), results and discussion about annotation styles (Section 5.2) and experiments with image slices from 3D data (Section 5.3), in this section we also present ways of improving meta-learning algorithms. As discussed in Section 4, we employed SSL pretraining to aid meta-learners in achieving better results in OOD tasks. More specifically, we show results with backbones pretrained with SimSiam (+**SS**) [29], more specifically, the ResNet-18 (**R18**) and ResNet-50 (**R50**).

Tables 4 and 5 present these results for *MSD-spleen* and *StructSeg-lungs*, respectively; using simply SSL pretraining with standard fine-tuning (Baseline), ANIL, PANet, R2D2 and MetaOptNet. As additional baseline comparisons to the larger **R18** and **R50** backbones, we show the EfficientLab (**EffLab**) and DeepLabv3 (**DLabv3**) results.

All the best overall results in Table 4 are achieved by backbones trained with SSL pretraining (**R50+SS** and **R18+SS**), mainly with gradient-based methods. In Table 5, overall best results for 5- and 10-shot are achieved using R2D2 paired with **R50+SS**. Also for *StructSeg-lungs* (Table 5), MetaOptNet outperforms other algorithms in extremely label-scarce 1-shot scenarios, once with **R18+SS** in the *points* annotation style and once with **DLabv3** in *scribbles* annotations.

Including all meta-learning approaches in our experiments – ANIL, PANets, R2D2 and MetaOptNet – the best results for each method in the 1-, 5- and 10-shots of *points* and *scribbles* were achieved by **R18+SS** or **R50+SS** in 36 out of 48 experiments in both *MSD-spleen* and *StructSeg-lungs*. Even when +**SS** backbones lose by some amount to **EffLab** and **DLabv3**, it is commonly by a considerably small margin, reinforcing the performance gains brought by SSL pretraining. These results make such strategies more suitable for generalizing to novel unseen data in real-world applications of this research. We point, however, that the meta-training phase using **R50+SS** coupled with gradient- and metric-based methods presented some stability during training. This instability renders **R18+SS** a safer approach, possibly given the smaller number of trainable parameters that better suit the MAML-based head of ANIL. Also, the much larger computational resources required by **R50**, make **R18** a considerably cheaper alternative for backbone choice with SSL pretraining.

6. Conclusion

In this work we introduced a new method for generalizing meta-learners from three distinct paradigms (gradient-based [9–12], metric-based [13,14] and fusion-based [15–17]) for FWS tasks without requiring ImageNet pretraining or strong domain-dependent priors. Some relevant improvements to existing methods have proven to be important for the success of the proposed approach. We chose to apply our methodology to radiology images, even though the same methods should also apply to other non-RGB domains such as histopathology, remote sensing, seismic images, or even 1D temporal signals. The experimental results allowed for the proposition of guidelines to adopt the proposed methodology. A summary of our conclusions and observations can be seen in Table 6.

For gradient-based methods, we observed that performance peaks when applying low-cost second-order gradient-based algorithms (i.e. ANIL [12]) in the segmentation head. These methods are also more computationally effective than second-order optimization-based methods applied to the whole network (i.e. MAML [9] or MetaSGD [11]), which should allow them to be applied in scenarios that require inference on higher spatial resolution images or even in 3D radiology. Fully first-order methods (i.e. Reptile [10]), however, could not reach the same performance as ANIL, even if they are relatively quicker and more scalable to larger backbones. Future works in gradient-based methods might include alternative supervised loss functions (e.g. Dice [24] or Focal [25]).

In metric-based meta-learners, the cosine distance and prototype alignment regularization of PANets [14] proved to be more powerful than the simple Euclidean distance of ProtoNets [13,20]. Similarity-based methods, however, are only able to generalize to novel domains that are quite close to the meta-training domains/tasks, possibly due to the lack of global context in such models. Early experiments during this work tried to insert pixel location information in the distance computation with no visible effects on ProtoNets and PANets. Future research directions might be more successful in integrating local information with a global context in such a way that benefits segmentation tasks (e.g. via CRFs [38] or visual attention [39]).

Table 5
IoU results for improved meta-learners in the *StructSeg-lungs* dataset for *points* (*num_points = 5, radius = 3*) and *scribbles* (*proportion = 0.25, thickness = 4*). Bold results represent the best results for each algorithm across network architectures and underscored values indicate the best overall results in a whole column.

Method		*Points*			*Scribbles*		
		Shots			Shots		
		1	5	10	1	5	10
Baseline	**R18+SS**	.222	.396	.475	.178	**.224**	**.224**
	R50+SS	**.247**	**.447**	**.643**	**.304**	.191	.169
ANIL	**DLabv3**	**.435**	**.518**	**.540**	.481	.511	.520
	R18+SS	.396	.493	.514	.527	.588	.595
	R50+SS	.309	.398	.413	**.445**	**.548**	**.581**
PANet	**EffLab**	.348	.579	.568	**.209**	**.167**	**.167**
	R18+SS	**.480**	**.662**	.647	.167	.165	.166
	R50+SS	.359	.641	**.676**	.168	.165	.165
R2D2	**DLabv3**	.338	.681	.718	.371	.381	.400
	R18+SS	.395	.694	.740	.360	.305	.327
	R50+SS	**.379**	**.755**	**.794**	**.508**	**.639**	**.680**
MetaOptNet	**DLabv3**	.395	**.714**	.763	**.588**	**.531**	**.610**
	R18+SS	**.462**	.695	.741	.412	.361	.394
	R50+SS	.371	.684	**.776**	.496	.506	.564

Table 6
Summary of conclusions on FWS meta-learners.

Annotation	Grid Skeletons & Scribbles	• Promising results, but closer to a costly dense annotation • More cost- and time-efficient annotations
Methods	Metric-based	• Easier implementation, straightforward convergence, but yields better results with supervised pretraining [12,14] • Works best in target tasks with small domain shifts
	Gradient- & Fusion-based	• Better for OOD tasks with unknown domain shifts • Fusion-based are less computationally expensive
Pre-training	SSL	• Improved results in OOD target tasks (Tables 4 and 5) • Considerable improvements for meta-learners in label-scarce scenarios (e.g. 1-shot/1-point in Tables 4 and 5) • Unstable for larger architectures (e.g. ResNet50)

Previous works on FWS, such as Guided Nets [17] and PANets [14], proved unable to adapt to novel domains with large domain shifts in relation to the meta-training datasets. In such scenarios, fusion-based approaches [15,16] were considered the optimal choices, followed by gradient-based methods [12]. These methods are also more computationally effective than second-order optimization-based methods applied to the whole network (e.g. MAML [9] or MetaSGD [11]), which should allow them to be applied in scenarios that require inference on higher spatial resolution images or even in 3D radiology.

Very challenging segmentation tasks that require context and texture analysis, with organs that do not have clearly defined borders (such as *STAP-cerebellum* [36]) are still quite hard to learn from in few-shot settings. In future works, we intend to port the methods presented in this paper to fully 3D data (CTs, MRIs, and PET scans) instead of reducing these volumes to 2D slices. We hope that the additional context from the 3rd dimension will aid the algorithms in learning these challenging 3D tasks, despite the higher computation cost involved in learning 3D convolutional kernels.

At last, another major limitation of our pipeline is the need for annotated data from related tasks, restricting the application of the Meta-Learning pretraining to domains wherein labeled data is available for multiple datasets. Aiming to mitigate this limitation, another promising direction might be to merge Meta-Learning with SSL pseudo labels in order to eliminate the need for annotated datasets from related domains.

CRediT authorship contribution statement

Hugo Oliveira: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Pedro H.T. Gama:** Validation, Software, Investigation, Conceptualization, Writing – original draft, Writing – review & editing. **Isabelle Bloch:** Funding acquisition, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing. **Roberto Marcondes Cesar Jr.:** Funding acquisition, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

One dataset is private (STAP/UC/USP), while all others are publicly available. All datasets are public and referenced in the manuscript, but we cannot redistribute them on our own.

Acknowledgments

The authors would like to thank FAPESP, Brazil (grants #2015/22308-2, #2017/50236-1, #2020/06744-5 and #2022/15304-4), Serrapilheira Institute, Brazil (grant #R-2011-37776), ANR, France (project #ANR-17-CE23-0021), CNPq, Brazil, CAPES, Brazil, FINEP, Brazil and MCTI, Brazil PPI-SOFTEX (TIC 13 DOU 01245.010222/2022-44) for their financial support for this research.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.patcog.2024.110471.

References

[1] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: NIPS, Vol. 25, 2012.

[2] J. Peng, Y. Wang, Medical image segmentation with limited supervision: A review of deep network models, IEEE Access 9 (2021) 36827–36851.

[3] H. Oliveira, R.M. Cesar, P.H. Gama, J.A. Dos Santos, Domain generalization in medical image segmentation via meta-learners, in: Conference on Graphics, Patterns and Images, Vol. 1, IEEE, 2022, pp. 288–293.

[4] P.H.T. Gama, H. Oliveira, J.A. dos Santos, R.M. Cesar Jr., An overview on meta-learning approaches for few-shot weakly-supervised segmentation, Comput. Graph. 113 (2023) 77–88.

[5] L. Jing, Y. Tian, Self-supervised visual feature learning with deep neural networks: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2020).

[6] T. Hospedales, A. Antoniou, P. Micaelli, A. Storkey, Meta-learning in neural networks: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 44 (9) (2021) 5149–5169.

[7] S. Luo, Y. Li, P. Gao, Y. Wang, S. Serikawa, Meta-seg: A survey of meta-learning for image segmentation, Pattern Recognit. 126 (2022) 108586.

[8] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, P. Yu, Generalizing to unseen domains: A survey on domain generalization, IEEE Trans. Knowl. Data Eng. 35 (8) (2022) 8052–8072.

[9] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: ICML, PMLR, 2017, pp. 1126–1135.

[10] A. Nichol, J. Schulman, Reptile: A scalable meta-learning algorithm, 2018, p. 4, arXiv preprint arXiv:1803.02999. 2 (3).

[11] Z. Li, F. Zhou, F. Chen, H. Li, Meta-SGD: Learning to learn quickly for few-shot learning, 2017, arXiv preprint arXiv:1707.09835.

[12] A. Raghu, M. Raghu, S. Bengio, O. Vinyals, Rapid learning or feature reuse? Towards understanding the effectiveness of MAML, 2019, arXiv preprint arXiv:1909.09157.

[13] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, in: NeurIPS, Vol. 30, 2017.

[14] K. Wang, J.H. Liew, Y. Zou, D. Zhou, J. Feng, PANet: Few-shot image semantic segmentation with prototype alignment, in: CVPR, 2019, pp. 9197–9206.

[15] K. Lee, S. Maji, A. Ravichandran, S. Soatto, Meta-learning with differentiable convex optimization, in: CVPR, 2019, pp. 10657–10665.

[16] L. Bertinetto, J.F. Henriques, P.H. Torr, A. Vedaldi, Meta-learning with differentiable closed-form solvers, in: ICLR, 2019.

[17] K. Rakelly, E. Shelhamer, T. Darrell, A.A. Efros, S. Levine, Few-shot segmentation propagation with guided networks, 2018, arXiv preprint arXiv:1806.07373.

[18] G. Cheng, C. Lang, J. Han, Holistic prototype activation for few-shot segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 45 (4) (2022) 4650–4666.

[19] C. Lang, G. Cheng, B. Tu, C. Li, J. Han, Base and meta: A new perspective on few-shot segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2023).

[20] P.H.T. Gama, H.N. Oliveira, J. Marcato, J. Dos Santos, Weakly supervised few-shot segmentation via meta-learning, IEEE Trans. Multimed. 25 (2022).

[21] S.M. Hendryx, A.B. Leach, P.D. Hein, C.T. Morrison, Meta-learning initializations for image segmentation, 2019, arXiv preprint arXiv:1912.06290.

[22] P.H. Gama, H. Oliveira, J.A. dos Santos, Learning to segment medical images from few-shot sparse labels, in: Conference on Graphics, Patterns and Images, IEEE, 2021, pp. 89–96.

[23] Z. Chang, Y. Lu, X. Ran, X. Gao, X. Wang, Few-shot semantic segmentation: A review on recent approaches, Neural Comput. Appl. 35 (25) (2023) 18251–18275.

[24] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: International Conference on 3D Vision, IEEE, 2016, pp. 565–571.

[25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: ICCV, 2017, pp. 2980–2988.

[26] X. Zhang, Y. Wei, Y. Yang, T.S. Huang, SG-one: Similarity guidance network for one-shot semantic segmentation, IEEE Trans. Cybern. 50 (9) (2020) 3855–3865.

[27] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: MICCAI, Springer, 2015, pp. 234–241.

[28] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, 2017, arXiv preprint arXiv:1706.05587.

[29] X. Chen, K. He, Exploring simple siamese representation learning, in: CVPR, 2021, pp. 15750–15758.

[30] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778.

[31] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.

[32] A.H. Abdi, S. Kasaei, M. Mehdizadeh, Automatic segmentation of mandible in panoramic X-Ray, J. Med. Imaging 2 (4) (2015) 044003.

[33] J. Rajendran, A. Irpan, E. Jang, Meta-learning requires meta-augmentation, in: NeurIPS, Vol. 33, 2020, pp. 5705–5715.

[34] A.L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B.A. Landman, G. Litjens, B. Menze, et al., A large annotated medical image dataset for the development and evaluation of segmentation algorithms, 2019, arXiv preprint arXiv:1902.09063.

[35] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B.A. Landman, G. Litjens, B. Menze, O. Ronneberger, R.M. Summers, et al., The medical segmentation decathlon, Nat. Commun. 13 (1) (2022) 1–13.

[36] H. Oliveira, L. Penteado, J.L. Maciel, S.F. Ferraciolli, M.S. Takahashi, I. Bloch, R.C. Junior, Automatic segmentation of posterior fossa structures in pediatric brain MRIs, in: Conference on Graphics, Patterns and Images, IEEE, 2021, pp. 121–128.

[37] T.M.H. Hsu, W.Y. Chen, C.-A. Hou, Y.-H.H. Tsai, Y.-R. Yeh, Y.-C.F. Wang, Unsupervised domain adaptation with imbalanced cross-domain data, in: ICCV, 2015, pp. 4121–4129.

[38] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, P.H. Torr, Conditional random fields as recurrent neural networks, in: ICCV, 2015, pp. 1529–1537.

[39] T. Hu, P. Yang, C. Zhang, G. Yu, Y. Mu, C.G. Snoek, Attention-based multi-context guiding for few-shot semantic segmentation, in: AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 8441–8448.

Hugo Oliveira is a Computer Science professor at Universidade Federal de Viçosa, with B.Sc. and M.Sc. degrees at Universidade Federal da Paraíba and Ph.D. at Universidade Federal de Minas Gerais. Research areas include Machine/Deep Learning, Few-Shot Learning, Meta-Learning, SelfSupervised Learning, Domain Adaptation, Generative Models and Open Set Recognition.

Pedro Gama is currently a Doctoral candidate at UFMG, with B.Sc. Computational and Applied Mathematics and MSc. degrees by UFMG. His research is focused on Deep Learning and Computer Vision, specifically in Semantic Segmentation, although his interests include General Machine Learning, Pattern Recognition, Few-Shot Learning, Meta-Learning, and Remote Sensing.

Isabelle Bloch received the Ph.D. degree from Télécom Paris. She has been a Professor at Télécom Paris until 2020 and is now a Professor at Sorbonne Université. Her current research interests include mathematical morphology, fuzzy set theory, graph-based, knowledge-based object recognition, and medical imaging.

Roberto M. Cesar Jr is a full-professor of the USP. Graduated in Computer Science from IBILCE-UNESP, MS from UNICAMP and Ph.D. from USP and is a Researcher at InovaUSP and member of the Coordination at FAPESP. Experience in computer science, computer vision, pattern recognition, signal and image processing.