Transductive few-shot learning

by

Malik BOUDIAF

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE IN PARTIAL FULFILLMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, APRIL 26, 2023

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Ismail Ben Ayed, Thesis supervisor
Department of System Engineering, École de technologie supérieure

Mr. Pablo Piantanida, Thesis Co-Supervisor
International Laboratory on Learning Systems (ILLS), CNRS-Université Paris-Saclay

Mr. Rafael Menelau Cruz, Chair, Board of Examiners
Department of Software and Information Technology Engineering, École de technologie supérieure

Mr. Matthew Toews, Member of the Jury
Department of System Engineering, École de technologie supérieure

Ms. Olga Veksler, External Examiner
Cheriton School of Computer Science, The University of Waterloo

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON APRIL 21, 2023

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

As this chapter closes, I would like to thank my thesis supervisor Ismail for these four incredible years of collaboration, growth, and friendship. For his generosity and dedication, I will always be grateful. I would also like to thank all my lab mates who've shared this long and enduring journey with me: Pierre, Hoël, Jérôme, Florent, Ziko and Saypra. Special thanks to Djebril and Soukaina, the friends who've become family and were always there to make me laugh (and feed me), in good and difficult times.

Second, I would like to thank my family. Although life took me far away from home, I've always kept each and every one of you on my mind and in my heart. Starting with my mother and father for their unconditional support throughout my studies, for being exceptionally loving and caring, and for the confidence they instilled in me. My brother Reda and my sisters Hana and Lina for being inspiring life models I'll always gravitate towards. And of course, my family-in-law for genuinely welcoming me as a member of their own.

Finally, I'd love to thank Sarah for her unconditional love and support. These years surely were stressful and testing at times, but you were always there to bring perspective back into my life.

Apprentissage *few-shot* par transduction

Malik BOUDIAF

RÉSUMÉ

Les modèles d'apprentissage profond ont connu un succès sans précédent, atteignant des performances proches de celles des humains lorsqu'ils sont entraînés sur des données étiquetées à grande échelle. Cependant, la capacité de généralisation de ces modèles peut être sérieusement remise en question lorsqu'il s'agit de traiter de nouvelles classes (non vues), avec seulement quelques instances étiquetées par classe. Les humains, en revanche, peuvent apprendre de nouvelles tâches rapidement à partir d'une poignée d'exemples, en exploitant le contexte et les connaissances préalables. Pour combler cet écart, la communauté en apprentissage automatique a développé au fil des années, des stratégies de méta-entraînement, dans le but de doter le modèle de capacités de généralisation intrinsèques.

Dans cette thèse, nous abordons le problème de l'apprentissage en quelques exemples sous un angle différent. Exploitant les opportunités qui émergent des *modèles de fondation*, ces grands modèles pré-entraînés une fois sur des ensembles de données comprenant des milliards d'exemples, nous transitionnons d'un paradigme centré sur l'entraînement à un paradigme centré sur l'inférence. Au travers de cette thèse, notre objectif est de développer des procédures d'inférence modulaires qui peuvent adapter efficacement n'importe quel modèle, indépendamment de son architecture ou de sa méthode d'entraînement, à des tâches d'apprentissage avec quelques exemples seulement. Pour accomplir cette tâche difficile, nous explorons les avantages et les limites de la transduction en tant que principe d'inférence, démontrant ainsi des résultats prometteurs sur des tâches de classification et de segmentation en quelques exemples.

En tant que première contribution, nous abordons la tâche courante de classification d'images en quelques exemples. Nous développons une procédure d'inférence transductive hautement modulaire, basée sur la maximisation de l'information mutuelle entre les caractéristiques extraites et les prédictions d'étiquettes. Nous observons des résultats très prometteurs, tant sur les benchmark expérimentaux usuels de l'apprentissage en quelques exemples que sur les benchmark présentant des écarts de domaine.

En tant que seconde contribution, nous explorons l'impact sur les méthodes transductives de l'introduction d'un déséquilibre de classes dans les données de test non étiquetées de chaque tâche. Nos résultats démontrent de forts effets indésirables pour toutes les méthodes transductives, conduisant certaines à sous-performer par rapport aux méthodes inductives de référence. Pour faire face à ce problème, nous diagnostiquons et étendons la procédure d'inférence basée sur l'information mutuelle décrite précédemment avec des divergences α, dont les gradients permettent une plus grande déviation de la distribution uniforme codée dans l'information mutuelle. Sur le plan empirique, nous observons des gains substantiels dans le scénario de déséquilibre de classes.

En tant que troisième contribution, nous continuons à explorer les propriétés potentiellement nuisibles des données non étiquetées sur les méthodes transductives. En particulier, nous étudions le problème d'*open-set*, dans lequel des classes perturbatrices peuvent être introduites dans les données non étiquetées. Motivés par l'observation que les méthodes transductives existantes présentent de mauvaises performances dans les scénarios d'*open-set*, nous proposons une généralisation du principe du maximum de vraisemblance, dans laquelle des scores latents réduisant l'influence des valeurs aberrantes potentielles sont introduits aux côtés du modèle paramétrique habituel. Nous montrons que cette méthode surpasse les méthodes inductives et transductives existantes sur les deux aspects de la reconnaissance *open-set*, à savoir la classification et la détection des valeurs aberrantes.

En guise de contribution finale, nous nous penchons sur la tâche difficile de la segmentation en quelques exemples, qui se caractérise par la présence combinée de tous les effets néfastes mentionnés ci-dessus: déséquilibre de classes et *open-set*. Nous présentons la première méthode qui abandonne complètement le méta-apprentissage et les architectures customisées. A la place, notre méthode utilise un modèle profond standard, entraîné par entropie croisée, et se concentre sur la formulation d'une inférence transductive par image pour chaque nouvelle tâche. Au-delà de la simplicité, nous trouvons que cette nouvelle approche de la segmentation en quelques exemples présente de forts avantages, notamment une capacité considérablement améliorée à exploiter une quantité croissante de supervision, dépassant de 6 % le précédent état de l'art en mIoU dans le scénario à 10 exemples, sur le benchmark le plus populaire.

Mots-clés: apprentissage few-shot, classification, segmentation sémantique, transduction

Transductive few-shot learning

Malik BOUDIAF

ABSTRACT

Deep learning models have achieved unprecedented success, approaching human-level performances when trained on large-scale labeled data. However, the generalization of such models might be seriously challenged when dealing with new (unseen) classes, with only a few labeled instances per class. Humans, however, can learn new tasks rapidly from a handful of instances, by leveraging context and prior knowledge. To bridge this gap, the few-shot learning community has relied on meta-training strategies, in an attempt to provide the model with intrinsic generalization abilities.

In this thesis, we see the few-shot problem in a different light. Noticing the opportunities emerging from *foundation models*, those large pre-trained models training once on billion-scaled datasets, we shift from the usual *training*-centered paradigm to an *inference*-centered one. Throughout this thesis, we aim to develop modular inference procedures that can efficiently adapt any model, regardless of its architecture or how it was trained, to few-shot tasks. To achieve that challenging task, we explore the benefits and limitations of transduction as an inference principle, demonstrating promising results on few-shot classification and few-shot segmentation tasks.

As a first contribution, we tackle the most popular problem of few-shot image classification. We develop a highly modular, transductive inference procedure based on the maximization of the mutual information between extracted features and label predictions. We observe very promising results, in both standard few-shot settings, and with domain shift between labeled and unlabeled samples.

As a second contribution, we explore the impact on transductive methods of introducing class imbalance in the unlabeled test data of each task. Our findings demonstrate strong adverse effects for all transductive methods, leading some to underperform inductive baselines. To cope with that setting, we diagnose and extend the mutual information-based inference procedure previously described with α-divergences, whose gradients allow more deviation from the uniform prior encoded in the mutual information. Empirically, we observe substantial gains in the class-imbalanced scenario.

As a third contribution, we continue to explore potential adverse properties of the unlabeled data on transductive methods. In particular, we investigate the few-shot open-set problem, in which distracting classes can be introduced in the unlabeled data. Motivated by the observation that existing transductive methods perform poorly in open-set scenarios, we propose a generalization of the maximum likelihood principle, in which latent scores down-weighing the influence of potential outliers are introduced alongside the usual parametric model. We show that this method surpasses existing inductive and transductive methods on both aspects of open-set recognition, namely closed-set classification and outlier detection.

As a final contribution, we examine the challenging setting of few-shot segmentation, which exhibits both adverse effects mentioned above: class imbalance and openness. We present the first method to completely forego meta-learning and custom architectures. Instead, it uses a standard backbone, trained with standard cross-entropy, and focuses on formulating a per-image transductive inference for each new task. Beyond simplicity, we find this new approach exhibits strong advantages, including a much-improved capacity to leverage an increasing amount of supervision, surpassing by 6 % mIoU previous state-of-the-art in the 10-shot scenario, on the most popular few-shot benchmark.

Keywords: few-shot learning, classification, semantic segmentation, transduction

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

OSR	Open-Set Recognition
FSC	Few-Shot Classification
FSL	Few-Shot Learning
FSS	Few-Shot segmentation
FSOSR	Few-Shot Open-Set Recognition
NLP	Natural Language Processing
TIM	Transductive Information Maximization
MAML	Model-Agnostic Meta-Learning
OSLO	Open-Set Likelihood Optimization
RePRI	Region Proportion Regularized Inference
ADMM	Alternate Direction Method of Multipliers
GD	Gradient Descent
KKT	Karush-Kuhn-Tucker
i.i.d.	Independent and Identically Distributed
SVM	Support Vector Machine
k-NN	k-Nearest Neighbor
DNN	Deep Neural Networks
LSTM	Long Short-Term Memory
MLP	Multi-Layer Perceptron

ViT	Vision Transformer
RN	ResNet
WRN	Wide ResNet
VGG	Visual Geometry Group
MI	Mutual Information
CE	Cross-Entropy
KL	Kullback-Leibler
Acc	Accuracy
AUROC	Area Under the Receiver Operating Characteristic curve
mIoU	mean Intersection over Union
AUPR	Area Under the Precision-Recall curve
GPU	Graphical Processing Unit
FPS	Frames Per Second

LIST OF SYMBOLS AND UNITS OF MEASUREMENTS

$\mathcal{X}$	Space of images
$\mathcal{Y}$	Space of labels
$\boldsymbol{x}$	Input image
$\boldsymbol{y}$	One-hot label
$\boldsymbol{\phi}$	Parameters of a feature extractor
$\boldsymbol{\theta}$	Parameters of a classification head
$\mathcal{H}(\boldsymbol{p})$	Shannon entropy of distribution $\boldsymbol{p}$
$\mathcal{I}(\boldsymbol{p};\boldsymbol{q})$	Mutual information between $\boldsymbol{p}$ and $\boldsymbol{q}$
$\mathcal{D}_{\mathrm{KL}}$	Kullback-Leibler divergence
$\mathbb{S}$	Support set
$\mathbb{Q}$	Query set

INTRODUCTION

For long, deep learning has remained task-centered. In other words, in order to address some particular task, one would collect and annotate a medium to large-scale dataset, and use it to train a model for this particular application. While intuitive, this approach scales poorly on different fronts.

Limitations of the supervised paradigm. It has become evident from the last decade of research that scaling up models and data remain by far the leading drivers of both "in-distribution" performance and "out-of-distribution' robustness (Taori *et al.*, 2020). Unfortunately, large-scale data collection can be impractical, or even intractable. For instance, obtaining scans from rare tumors at scale is, by definition, impossible. Additionally, the annotation process can be extremely expensive, especially for those applications that require cutting-edge expertise. But even abstracting away the data collection problem previously mentioned, the computational cost of training state-of-the-art architectures is becoming a limiting factor for most research groups and companies. More than ever, the ML community needs to rethink standard pipelines in a way that better factorizes heavy, compute/energy-intensive training procedures.

Foundation models: new opportunities. "Foundation models" (Bommasani *et al.*, 2021) are emerging as an appealing framework for researchers and practitioners. In that framework, large models are "(pre)trained once" on vast, unlabeled, and weakly curated datasets that usually stand at the billion-samples scale. Notorious examples of such models first appeared in the Natural Language Processing (NLP) community, such as the popular GPT-3 (Brown *et al.*, 2020), and have recently landed in vision, with CLIP (Radford *et al.*, 2021), ALIGN (Jia *et al.*, 2021) among others. The idea is that once trained, these models can provide discriminative features that allow them to be adapted ad-hoc to a wide range of downstream scenarios, from a very limited amount of supervision.

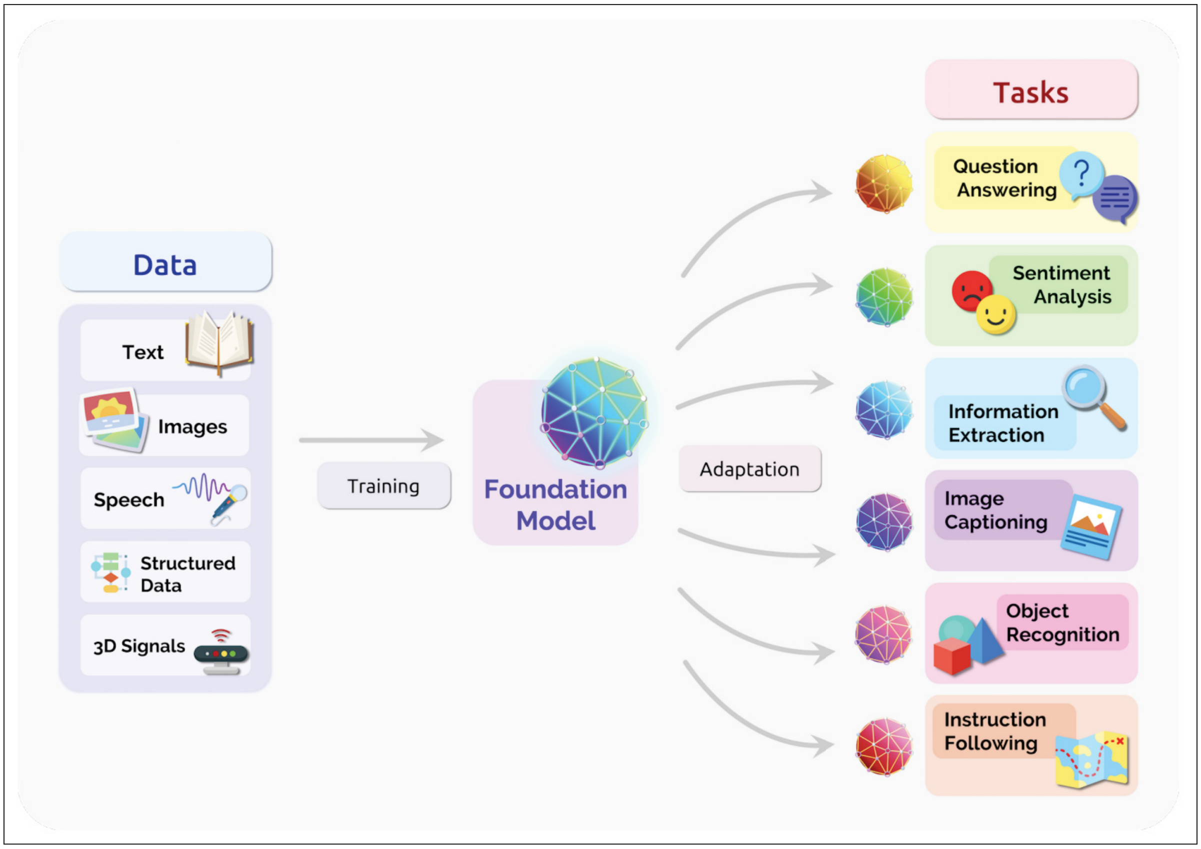

Figure 0.1 Figure taken from (Bommasani *et al.*, 2021), showing the overall "foundation models" framework. Vast, non-curated data is used to train a model once. This same model is then adapted to a variety of tasks using significantly fewer data and resources

Such a framework offers the potential to largely alleviate the data collection and annotation problems previously described. By factorizing the long, expensive, and highly energy-intensive large-scale pretraining process, this framework also enables the low-resource reuse of large models. As a by-product, this framework also greatly enhances reproducibility, standardizing architectures and pre-trained checkpoints, thereby limiting the potential use of bells and whistles when comparing methods.

Towards model-agnosticity. How to adapt a model to a particular task from limited supervision is far from being a trivial question. In fact, this question has fueled a whole research area, namely Few-shot learning (FSL). In FSL, the pre-trained model is tested by its ability to generalize from only a few labeled examples, in principle belonging to classes that the model has never seen before. Interestingly, while the FSL problem quite perfectly lends itself to the "foundation models" framework, the pretraining procedure, as well as customized architectures, have for long remained at the heart of few-shot methods. Only very recently, around the same year this thesis work was initiated, did a branch of the FSL literature start to approach the problem with a more "model-agnostic" perspective, abstracting away the pretraining procedure and the model architecture, and instead focusing on developing optimization-based inference procedures. Works presented in this thesis follow this track, and particularly focus on two tasks of interest: classification and semantic segmentation for natural images. With that said, all methods described throughout the thesis were developed in a task-agnostic spirit, such that they could trivially extend to further problems, even beyond computer vision.

Transduction as a promising avenue. Inductive learning remains the most common learning principle in computer vision, and consists in inferring general rules, usually in the form of a fitted parametric model, from training samples. These general rules are then applied to infer the label of unlabeled instances. Because the few-shot learning framework is characterized by extremely scarce supervision, the problem is highly ambiguous, such that an infinity of general rules could explain the labeled samples, leading to poor generalization. In this context, transduction becomes a particularly appealing learning principle. In Vapnik's words: *When solving a problem of interest, do not solve a more general problem as an intermediate step. Try to get the answer that you really need but not a more general one.* In essence, transduction advocates finding rules that work for the specific unlabeled test instances observed, and thereby bypasses the *induction* → *deduction* steps, as illustrated in Figure 0.2. By both assuming access to more data, e.g. unlabeled test instances, and lowering the target goal, e.g. finding a local rule rather than a general one, we show in this thesis that transductive learning is a promising, and practically interesting principle with a large upside potential.

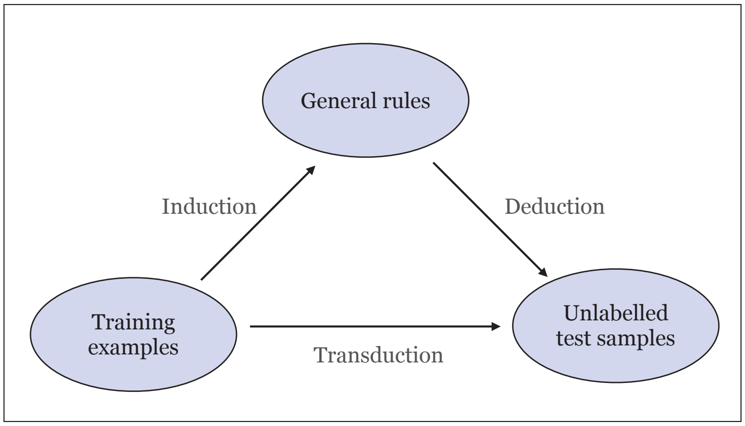

Figure 0.2 The widespread learning framework splits into two reasoning steps; the inductive step infers general rules from training examples, while the deductive step applies those rules to infer the label of test instances. Transduction does not attempt to find general rules and is only interested in providing predictions on the provided test samples

Research statement and challenges

In this thesis, we investigate ways to apply transductive learning principles to develop data-efficient and highly modular methods, applicable on top of virtually any existing pre-trained model, for the few-shot classification and segmentation problems. In particular, we present a nuanced view of transductive learning, exploring its upsides as well as its blind spots, and offering practical solutions. We summarize the three important research challenges that this thesis seeks to address as follows:

- **Efficiency** The general goal of few-shot methods learning is to develop models and inference procedures that are data-efficient. Although less of a formal requirement, compute efficiency is highly suitable for inference procedures, which are generally expected to run on a single GPU. By moving complexity from the training stage to the inference stage, we generally augment compute requirement of the latter. Therefore, the first challenge is to keep inference procedures lightweight enough so that they can tractably scale up to larger tasks, e.g. with more samples and classes.

- **Modularity** As we wish to develop inference procedures that can be effortlessly plugged into any existing model. Therefore, we must build methods that do not rely on particular assumptions about the properties of the feature distribution, or the specific architecture used. For instance, power transforms that require strictly positive features could not be used as a model-agnostic feature transformation.
- **Robustness** As we'll see throughout this thesis, transductive methods can be highly affected by adverse properties of the unlabeled data they seek to leverage. Therefore, an important challenge is to formulate inference procedures that can cope with such adverse properties, and ideally never fall under an inductive baseline.

Contributions

As formulated in the previous paragraph, we explore ways to leverage transductive learning to obtain model-agnostic inference procedures that perform well in the challenging few-shot setting.

The core contributions of this thesis are:

- In Chapter 2, we tackle the standard problem of few-shot closed-set image classification. We develop a highly modular, transductive inference procedure based on the maximization of the mutual information between extracted features and label predictions.

  Related publication:

  Information maximization for few-shot learning, published in *Neural Information Processing Systems (NeurIPS)*, 2020.
- In Chapter 3, we extend the standard few-shot classification setting. Specifically, we explore the impact of introducing Dirichlet-based class imbalance in the unlabeled test

data. We quantify the extent to which class imbalance adversely impacts the performance of transductive methods, and propose an extension of the method presented in Chapter 2 based on α-divergences.

Related publications:

Realistic Evaluation of Transductive Few-Shot Learning, published in *Neural Information Processing Systems (NeurIPS)*, 2021.

- In Chapter 4, we investigate another orthogonal extension of the standard few-shot classification setting. Specifically, we measure the influence of introducing outliers from distracting classes in the unlabeled test data commonly referred to as the (*open-set* scenario), and propose an effective modification of the traditional Maximum Likelihood Estimation principle that models the potential presence of outliers.

  Related publication:

  Open-Set Likelihood Maximization for Few-Shot Learning (accepted at the *Computer Vision and Pattern Recognition (CVPR)*, 2023)

- Finally, in Chapter 5, we explore the related but challenging few-shot segmentation setting. Segmentation is both class-imbalanced and open-set by nature, which makes it a very interesting setting to study for transductive learning. In this chapter, we present the first model-agnostic transductive inference procedure for the segmentation task.

  Related publication:

  Few-Shot Segmentation Without Meta-Learning: A Good Transductive Inference Is All You Need?, published in the *Computer Vision and Pattern Recognition (CVPR)*, 2021.

To facilitate further research and improve the reproducibility of results, all codes of the papers above are public.

Additional contributions

Additional contributions were made during the course of this thesis on various topics that we list below:

– *Information theoretic tools for training neural networks.* Throughout these two works, we focus on understanding and proposing loss functions for training Deep Neural Networks (DNNs) that lead to better generalization.

  Related publication:

  A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses, published in the *European Conference on Computer Vision (ECCV)*, 2020 - first author.

  A differential entropy estimator for training neural networks, published in the *International Conference on Machine Learning (ICML)*, 2022 – co-author.

– *Test-time adaptation.* In this work, we develop a per-batch transductive inference that obtains competitive results in the setting of online test-time adaptation, while being faster and simpler to tune.

  Related publication:

  Parameter-free Online Test-time Adaptation, published in the *Computer Vision and Pattern Recognition (CVPR)*, 2022 – first author.

– *Adversarial robustness.* In this work, we propose a new adversarial defense based on the Fisher-Rao regularization that obtains competitive clean, and robust performances, while significantly reducing the training time.

  Related publication:

  Adversarial Robustness via Fisher-Rao Regularization, published in *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2022 – co-author.

- *Extended few-shot settings*. These two works come in line with the works presented in this thesis, and explore more general and challenging settings for the few-shot classification and segmentation.

  Related publication:

  Towards Practical Few-Shot Query Sets: Transductive Minimum Description Length Inference, published in *Neural Information Processing Systems (NeurIPS)*, 2022 - co-author.
  A Strong Baseline for Generalized Few-Shot Semantic Segmentation, published in the *Computer Vision and Pattern Recognition (CVPR)*, 2023 - co-author.

CHAPTER 1

LITERATURE REVIEW

As mentioned in Introduction, the few-shot learning problem fits quite nicely the narrative of foundation models, with the *training* and *task-adaptation* stages clearly separated. In that scenario, the *training* stage is abstracted away, and a few-shot method is in charge of specifying an *adaptation* procedure that is as agnostic as possible to the architecture of the model, or the way it was trained.

While an appealing framework, mainstream few-shot approaches have, and predominantly still, entangle *training* AND *adaptation*, making both stages integral parts of a few-shot method. That entanglement comes as a direct consequence of the meta-learning paradigm, upon which most part of the FSL literature relies.

In this chapter, we start in section 1.1 by reviewing the mainstream meta-learning approaches for few-shot learning. That will naturally lead us to discuss trade-offs and limitations. We then proceed in section 1.2 with the more recent *model-agnostic* line of works that fit the *foundation models* narrative. Finally, in Section 1.3, we elaborate on the opportunities and challenges brought by transductive learning.

1.1 Meta-learning

As students, we're often told that the benefit of school is not the absolute knowledge we acquire throughout our curriculum, but rather our capacity to adapt and learn more rapidly in new situations. In other words, school makes us learn how to learn. What if we could apply the same to automated agents? In other words, how to teach a model to learn faster, and to be more data-efficient when presented with a new task. Those are the core research questions behind meta-learning. Formulated in this way, meta-learning's narrative sounds particularly promising for the few-shot learning problem. For instance, concrete instances of this paradigm could be finding a way for a model to decide its optimal learning hyperparameters, based on newly observed data.

1.1.1 Deep adaptation

That line of work aims to find the model's weights that lead to a fast and efficient finetuning of the same model, once presented with a new task. *MAML* (Finn *et al.*, 2017) remains by far the most popular work in that category. The objective is to train the model such that the final weights obtained can serve as a good initialization for any downstream task. At training time, *MAML* formulates a bi-level optimization problem, in which the outer optimization problem tries to adjust the model's weights such that few-gradient steps on a few labeled samples (inner problem) allow obtaining good generalization for this task. In practice, unrolling optimization steps on the whole model is computationally very intensive, such that bells and whistles turn out quite important in the final performance (Antoniou, Edwards & Storkey, 2019).

1.1.2 Shallow adaptation

Adapting all the weights of the model makes the bi-level optimization problem cumbersome. Instead, cheaper meta-learning alternatives have been proposed. While their formulation differ, they share a common goal of meta-learning a robust embedding that facilitates learning a shallow model for each task. Due to their proximity with metric-learning techniques, those methods are sometimes grouped under the denomination "metric-based" meta-learning. Popular methods in that category include *ProtoNet* (Snell, Swersky & Zemel, 2017), which meta-learns a simple centroid-based classifier, *MatchingNet* (Vinyals, Blundell, Lillicrap, Wierstra *et al.*, 2016b) which essentially kernelizes *ProtoNet*, *RelationNet* (Sung *et al.*, 2018) which meta-learns the pairwise similarity function altogether, or *MetaOpt* (Lee, Maji, Ravichandran & Soatto, 2019b) which meta-learns an SVM classifier.

Note that metric-learning based meta-learning has been also very popular for the few-shot segmentation setting. Particularly, the support images are employed to generate class prototypes, which are later used to segment the query images via a prototype-query comparison module. Early frameworks followed a dual-branch architecture, with two independent branches (Shaban, Bansal, Liu, Essa & Boots, 2018; Dong & Xing, 2018; Rakelly, Shelhamer, Darrell, Efros & Levine, 2018), one generating the prototypes from the support images and the other segmenting the query images with the learned prototypes. More recently, the dual-branch setting has been unified into a single-branch architecture, which employs the same embedding function for both the support and query sets (Zhang, Wei, Yang & Huang, 2020b; Siam, Oreshkin & Jagersand, 2019; Wang, Liew, Zou, Zhou & Feng, 2019a; Yang, Liu, Li, Jiao & Ye, 2020a; Liu, Zhang, Zhang & He, 2020f). These approaches mainly aim at exploiting better guidance for the segmentation of query images (Zhang *et al.*, 2020b; Nguyen & Todorovic, 2019; Wang *et al.*, 2020a; Zhang *et al.*, 2019a), by learning better class-specific representations (Wang *et al.*, 2019a; Liu, Zhang, Lin & Liu, 2020d; Liu *et al.*, 2020f; Yang *et al.*, 2020a; Siam *et al.*, 2019) which are later used to segment the query images via a prototype query comparison module.

1.1.3 Limitations

Although a promising paradigm, meta-learning has shown several limitations over the years that have gradually come to question its superiority as a learning paradigm for FSL. We summarize them in three important points:

- **Performances.** Recent evidence casts large doubts on the benefit of meta-learning, at least as has been applied so far. The first interesting results on the matter came from (Chen, Liu, Kira, Wang & Huang, 2019), who showed through extensive experiments that fairly reproduced methods did not compare favorably to a simple baseline using a standard cross entropy-based training followed by a per-task linear classifier at test-time. In fact, authors show that when training/testing conditions do not match, e.g. the number of classes used to simulate training tasks does not correspond to the actual number of ways observed at test time, meta-learning approaches tend to be largely outperformed by simple baseline. (Cao, Law & Fidler, 2020) formalizes a saturation phenomenon when training and testing shots differ. We provide additional empirical evidence for this phenomenon in (Boudiaf *et al.*, 2021), presented in Chapter 5. Interestingly, (Laenen & Bertinetto, 2021) demonstrates the sub-optimality of episodic training for *ProtoNet*, showing that meta-learning a prototypical classifier practically results in using fewer negative pairs, and therefore less supervising signal than a standard metric-learning objective.

- **Scalability.** Besides questionable performance benefits, the episodic (or *multitask*) training, as well as the bi-level optimization procedure upon which meta-learning methods rely makes it challenging to distribute and upscale those methods. For instance, training *MAML* on modern state-of-the-art models, such as large vision transformers, would be a full-fledged engineering challenge.
- **Modularity.** Finally, by relying on customized training procedures, and sometimes customized architectures, meta-learning approaches can neither be seamlessly integrated into existing pipelines and be applied on top of open-source off-the-shelf models. At a time when foundation models are taking over, with spectacular performances over a wide range of tasks, the ability to fully benefit from the latest advances in image recognition becomes an increasingly important asset for few-shot methods, but one that meta-learning-based methods do not possess.

1.2 Model-agnostic approaches

A recent line of work has emerged for few-shot classification that has the potential to solve the three limitations listed above. Although not formally named in the literature, we will call this branch *model-agnostic approaches*. The idea is to disentangle the training phase from the test-time task adaptation, and take as few assumptions as possible regarding the architecture of the provided model, or how it was trained.

Inductive baselines. Earlier works in that category laid down strong inductive baselines. To the best of our knowledge, (Chen *et al.*, 2019) was the first work to propose an exhaustive and fair study of existing methods, and compare them to a naive baseline that trains a model with a vanilla cross-entropy. At test time, this baseline infers the class of unlabeled instances using the cosine similarity with labeled points in the feature space of the frozen model at test time. The baseline was shown to match existing methods in common scenarios and significantly outperform them in “extended” scenarios, including for instance more classes than commonly used, or a domain shift between labeled and unlabeled instances of a task. (Wang, Chao, Weinberger & van der Maaten, 2019b) adds up to this evidence by showing that a simple centering transformation applied before taking the cosine similarity systematically boosts the results, and is the first to show results across six different model architectures. Interestingly, (Tian, Wang, Krishnan, Tenenbaum & Isola, 2020a) show that more advanced training techniques developed for standard image classification, such as model distillation, also positively impacted the few-shot performance of such baselines, which goes on to demonstrate the importance of the modularity property of few-shot methods.

Transductive methods. As mentioned in Introduction, induction/deduction are natural reasoning steps, and by far the most popular in Computer Vision. Nevertheless, their relevance remains limited in under-specified settings such as the few-shot one. On the other hand, transduction can disambiguate the problem, and provide significant performance gains in settings where data is scarce (Vapnik, 2013), and has recently emerged as a whole branch of the few-shot literature, showing staggering performances compared to their inductive counterparts. Such works generally rely on enforcing the "cluster assumption". To the best of our knowledge, (Liu *et al.*, 2019c) was the first to apply transduction to a few-shot learning problem, in order to propagate labels to unlabeled data. (Liu, Song & Qin, 2020c) proposes a pseudo-labeling scheme to rectify prototypes, akin to the first iterations of a K-means algorithm, while (Ziko, Dolz, Granger & Ayed, 2020) explicitly formulates each task inference as a clustering problem, in which a Laplacian regularization is applied to unlabeled samples. (Dhillon, Chaudhari, Ravichandran & Soatto, 2020) and (Boudiaf *et al.*, 2020a) (presented in Chapter 2 of this thesis), develop entropy and mutual-information-based losses to perform clustering. Finally, (Hu, Gripon & Pateux, 2021) cast the problem of assigning labels to unlabeled samples as an optimal transport problem, which they solve through a Sinkhorn algorithm (Cuturi, 2013).

1.3 Transduction: opportunities and limitations

While a transductive approach appears particularly promising to the few-shot problem, even more so when no assumptions are to be taken during the training phase, transduction is no *free-lunch*. By making joint predictions on labeled and unlabeled data, transduction presupposes the simultaneous availability of test instances at inference, which arguably may not hold for all applications.

More importantly, leveraging unlabeled data without any prior about it can be challenging. Some properties of an underlying data distribution are known to negatively affect learning, even in the standard large-scale supervised settings. However, we argue that such properties become all the more problematic when they characterize the unlabeled data distribution that we aim to leverage, for the very reason that we cannot even be aware of them. For instance, (Lichtenstein, Sattigeri, Feris, Giryes & Karlinsky, 2020) first presented results on the strongly adverse effect of introducing noise, in the form of distracting classes in the unlabeled data of each task. Although preliminary, (Hu *et al.*, 2021) showed that unevenly balanced class distribution in the query set could also result in dramatic performance degradation.

Understanding, measuring, and mitigating such adverse effects are the very subject of Chapter 4. Chapter 5 also implicitly addresses these questions by tackling the semantic segmentation problem, which is naturally open-set due to the presence of distracting background and class imbalance.

Links to classical computer vision. To finish this section of related works, we wish to draw a parallel between the line of research presented throughout this thesis and *classical* (or *pre deep learning*) computer vision research. Transduction is not a novel concept, and informally speaking, trades off general learning for ad-hoc problem-solving. Therefore, although not explicitly termed as such, this concept was largely employed in classical computer vision for tasks that were either too hard to solve using general rules, either because the scale of the data did not allow to draw such rules reliably, or because we did not yet possess tractable ways to learn expressive enough models. As an illustrative example, we can think of popular graph-cut (Boykov & Funka-Lea, 2006) techniques used to address segmentation tasks. Considering all pixels from a given image as observed, unlabeled variables, a graph-cut inference works by formulating a latent-variable problem for each new image encountered, and developing efficient solvers. Initializing the variables of this problem may be done inductively, i.e. we may have a rough prior of the pixel color distribution of a banana, but the heavy lifting is still done ad-hoc, after observing the actual pixels of a specific image. Conceptually, the methods developed in this thesis only differ from this type of approach by their use of higher-dimensional learned features in place of raw (or manually crafted) ones, and by the tools used to formulate optimization objectives and solve them.

CHAPTER 2

TRANSDUCTIVE INFORMATION MAXIMIZATION FOR FEW-SHOT LEARNING

Malik Boudiaf[1] , Imtiaz Masud Ziko[1] , Jérôme Rony[1] , Jose Dolz[1] , Pablo Piantanida[2] ,
Ismail Ben Ayed[1]

[1] ÉTS Montréal, QC, Canada,
[2] Laboratoire des Signaux et Systèmes (L2S),
CentraleSupelec-CNRS-Université Paris-Saclay, France

Poster at Neural Information Processing Systems (NeurIPS) 2020, Vancouver.

Abstract

We introduce Transductive Infomation Maximization (TIM) for few-shot learning. Our method maximizes the mutual information between the query features and their label predictions for a given few-shot task, in conjunction with a supervision loss based on the support set. Furthermore, we propose a new alternating-direction solver for our mutual-information loss, which substantially speeds up transductive-inference convergence over gradient-based optimization, while yielding similar accuracy. TIM inference is modular: it can be used on top of any base-training feature extractor. Following standard transductive few-shot settings, our comprehensive experiments[1] demonstrate that TIM outperforms state-of-the-art methods significantly across various datasets and networks, while used on top of a fixed feature extractor trained with simple cross-entropy on the base classes, without resorting to complex meta-learning schemes. It consistently brings between 2% and 5% improvement in accuracy over the best performing method, not only on all the well-established few-shot benchmarks but also on more challenging scenarios, with domain shifts and larger numbers of classes.

[1] Code publicly available at https://github.com/mboudiaf/TIM

2.1 Introduction

Deep learning models have achieved unprecedented success, approaching human-level performances when trained on large-scale labeled data. However, the generalization of such models might be seriously challenged when dealing with new (unseen) classes, with only a few labeled instances per class. Humans, however, can learn new tasks rapidly from a handful of instances, by leveraging context and *prior* knowledge. The few-shot learning (FSL) paradigm (Miller, Matsakis & Viola, 2000b; Fei-Fei, Fergus & Perona, 2006; Vinyals *et al.*, 2016b) attempts to bridge this gap, and has recently attracted substantial research interest, with a large body of very recent works, e.g., (Hou, Chang, Bingpeng, Shan & Chen, 2019; Dhillon *et al.*, 2020; Rusu *et al.*, 2019; Ye, Hu, Zhan & Sha, 2020a; Liu *et al.*, 2019c; Chen *et al.*, 2019; Qiao *et al.*, 2019; Kim, Kim, Kim & Yoo, 2019; Sung *et al.*, 2018; Wertheimer & Hariharan, 2019; Gidaris, Bursuc, Komodakis, Pérez & Cord, 2019; Snell *et al.*, 2017; Finn *et al.*, 2017), among many others. In the few-shot setting, a model is first trained on labeled data with *base* classes. Then, model generalization is evaluated on few-shot *tasks*, composed of unlabeled samples from novel classes unseen during training (the *query* set), assuming only one or a few labeled samples (the *support* set) are given per novel class.

Most of the existing approaches within the FSL framework are based on the "learning to learn" paradigm or meta-learning (Finn *et al.*, 2017; Snell *et al.*, 2017; Vinyals *et al.*, 2016b; Sung *et al.*, 2018; Lee *et al.*, 2019b), where the training set is viewed as a series of balanced tasks (or *episodes*), so as to simulate test-time scenario. Popular works include prototypical networks (Snell *et al.*, 2017), which describes each class with an embedding prototype and maximizes the log-probability of query samples via episodic training; matching network (Vinyals *et al.*, 2016b), which represents query predictions as linear combinations of support labels and employs episodic training along with memory architectures; MAML (Finn *et al.*, 2017), a meta-learner, which trains a model to make it "easy" to fine-tune; and the LSTM meta-learner in (Ravi & Larochelle, 2016), which suggests optimization as a model for few-shot learning. A large body of meta-learning works followed-up lately, to only cite a few (Rusu *et al.*, 2019;

Oreshkin, López & Lacoste, 2018; Mishra, Rohaninejad, Chen & Abbeel, 2018; Sung *et al.*, 2018; Ye *et al.*, 2020a).

2.1.1 Related work

Transductive inference: In a recent line of work, *transductive* inference has emerged as an appealing approach to tackling few-shot tasks (Dhillon *et al.*, 2020; Hou *et al.*, 2019; Kim *et al.*, 2019; Liu *et al.*, 2019c; Qiao *et al.*, 2019; Nichol, Achiam & Schulman, 2018; Liu, Song & Qin, 2019a; Ziko *et al.*, 2020), showing performance improvements over *inductive* inference. In the transductive setting[2], the model classifies the unlabeled query examples of a single few-shot task at once, instead of one sample at a time as in inductive methods. These recent experimental observations in few-shot learning are consistent with established facts in classical transductive inference (Vapnik, 1999; Joachims, 1999; Dengyong, Bousquet, Lal, Weston & Schölkopf, 2004), which is well-known to outperform inductive methods on small training sets. While (Nichol *et al.*, 2018) used information of unlabeled query samples via batch normalization, the authors of (Liu *et al.*, 2019c) were the first to model explicitly transductive inference in few-shot learning. Inspired by popular label-propagation concepts (Dengyong *et al.*, 2004), they built a meta-learning framework that learns to propagate labels from labeled to unlabeled instances via a graph. The meta-learning transductive method in (Hou *et al.*, 2019) used attention mechanisms to propagate labels to unlabeled query samples. More closely related to our work, the recent transductive inference of Dhillion et al. (Dhillon *et al.*, 2020) minimizes the entropy of the network softmax predictions at unlabeled query samples, reporting competitive few-shot performances, while using standard cross-entropy training on the base classes. The competitive performance of (Dhillon *et al.*, 2020) is in line with several recent inductive baselines (Chen *et al.*, 2019; Wang *et al.*, 2019b; Tian *et al.*, 2020a), which reported that standard cross-entropy training for the base classes matches or exceeds the performances of more sophisticated meta-learning procedures. Also, the performance of (Dhillon *et al.*, 2020)

[2] Transductive few-shot inference is not to be confused with semi-supervised few-shot learning (Ren *et al.*, 2018). The latter uses extra unlabeled data during meta-training. Transductive inference has access to exactly the same training/testing data as its inductive counterpart.

is in line with established results in the context of semi-supervised learning, where entropy minimization is widely used (Grandvalet & Bengio, 2005; Miyato, Maeda, Koyama & Ishii, 2018; Berthelot *et al.*, 2019). It is worth noting that the inference runtimes of transductive methods are, typically, much higher than their inductive counterparts. For, instance, the authors of (Dhillon *et al.*, 2020) fine-tune all the parameters of a deep network during inference, which is several orders of magnitude slower than inductive methods such as ProtoNet (Snell *et al.*, 2017). Also, based on matrix inversion, the transductive inference in (Liu *et al.*, 2019c) has a complexity that is cubic in the number of query samples.

Info-max principle: While the semi-supervised and few-shot learning works in (Grandvalet & Bengio, 2005; Dhillon *et al.*, 2020) build upon Barlow's principle of entropy minimization (Barlow, 1989), our few-shot formulation is inspired by the general info-max principle enunciated by Linsker (Linsker, 1988), which formally consists in maximizing the Mutual Information (MI) between the inputs and outputs of a system. In our case, the inputs are the query features and the outputs are their label predictions. The idea is also related to info-max in the context of clustering (Krause, Perona & Gomes, 2010; Hu, Miyato, Tokui, Matsumoto & Sugiyama, 2017; Jabi, Pedersoli, Mitiche & Ayed, 2019). More generally, info-max principles, well-established in the field of communications, were recently used in several deep-learning problems, e.g., representation learning (Hjelm *et al.*, 2019; Oord, Li & Vinyals, 2018), metric learning (Boudiaf *et al.*, 2020b) or domain adaptation (Liang, Hu & Feng, 2020), among other problems.

2.1.2 Contributions

- We propose Transductive Information Maximization (TIM) for few-shot learning. Our method maximizes the MI between the query features and their label predictions for a few-shot task at inference, while minimizing the cross-entropy loss on the support set.
- We derive an alternating-direction solver for our loss, which substantially speeds up transductive inference over gradient-based optimization, while yielding competitive accuracy.
- Following standard transductive few-shot settings, our comprehensive evaluations show that TIM outperforms state-of-the-art methods substantially across various datasets and

networks, while using a simple cross-entropy training on the base classes, without complex meta-learning schemes. It consistently brings between 2% and 5% of improvement in accuracy over the best performing method, not only on all the well-established few-shot benchmarks but also on more challenging, recently introduced scenarios, with domain shifts and larger numbers of ways. Interestingly, our MI loss includes a label-marginal regularizer, which has a significant effect: it brings substantial improvements in accuracy, while facilitating optimization, reducing transductive runtimes by orders of magnitude.

2.2 Transductive Information Maximization

2.2.1 Few-shot setting

Assume we are given a labeled training set, $\mathcal{X}_{\text{base}} := \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{N_{\text{base}}}$, where $\boldsymbol{x}_i$ denotes raw features of sample i and $\boldsymbol{y}_i$ its associated one-hot encoded label. Such labeled set is often referred to as the *meta-training* or *base* dataset in the few-shot literature. Let $\mathcal{Y}_{\text{base}}$ denote the set of classes for this base dataset. The few-shot scenario assumes that we are given a *test* dataset: $\mathcal{X}_{\text{test}} := \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{N_{\text{test}}}$, with a completely new set of classes $\mathcal{Y}_{\text{test}}$ such that $\mathcal{Y}_{\text{base}} \cap \mathcal{Y}_{\text{test}} = \emptyset$, from which we create randomly sampled few-shot *tasks*, each with a few labeled examples. Specifically, each K-way $N_{\mathbb{S}}$-shot task involves sampling $N_{\mathbb{S}}$ labeled examples from each of K different classes, also chosen at random. Let $\mathbb{S}$ denote the set of these labeled examples, referred to as the *support* set with size $|\mathbb{S}| = N_{\mathbb{S}} \cdot K$. Furthermore, each task has a *query* set denoted by $\mathbb{Q}$ composed of $|\mathbb{Q}| = N_{\mathbb{Q}} \cdot K$ unlabeled (unseen) examples from each of the K classes. With models trained on the base set, few-shot techniques use the labeled support sets to adapt to the tasks at hand, and are evaluated based on their performances on the unlabeled query sets.

2.2.2 Proposed formulation

We begin by introducing some basic notation and definitions before presenting our overall Transductive Information Maximization (TIM) loss and the different optimization strategies for tackling it. For a given few-shot task, with a support set $\mathbb{S}$ and a query set $\mathbb{Q}$, let X denote the random variable associated with the raw features within $\mathbb{S} \cup \mathbb{Q}$, and let $Y \in \mathcal{Y} = \{1, \ldots, K\}$ be the random variable associated with the labels. Let $f_{\boldsymbol{\phi}} : \mathcal{X} \longrightarrow \mathcal{Z} \subset \mathbb{R}^d$ denote the encoder (*i.e.*, feature-extractor) function of a deep neural network, where $\boldsymbol{\phi}$ denotes the trainable parameters, and $\mathcal{Z}$ stands for the set of embedded features. The encoder is first trained from the base training set $\mathcal{X}_{\text{base}}$ using the standard cross-entropy loss, without any meta training or specific sampling schemes. Then, for each specific few-shot task, we propose to minimize a mutual-information loss defined over the query samples.

Formally, we define a soft-classifier, parametrized by weight matrix $\boldsymbol{\theta} \in \mathbb{R}^{K \times d}$, whose posterior distribution over labels given features[3], $p_{ik} := \mathbb{P}(Y = k | X = \boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\phi})$, and marginal distribution over query labels, $\widehat{p}_k = \mathbb{P}(Y_{\mathbb{Q}} = k; \boldsymbol{\theta}, \boldsymbol{\phi})$, are given by:

$$p_{ik} \propto \exp\left(-\frac{\tau}{2} \|\boldsymbol{\theta}_k - \boldsymbol{z}_i\|^2\right), \quad \text{and} \quad \widehat{p}_k = \frac{1}{|\mathbb{Q}|} \sum_{i \in \mathbb{Q}} p_{ik} \tag{2.1}$$

where $\boldsymbol{\theta} := [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K]$ denotes classifier weights, $\boldsymbol{z}_i = \frac{f_{\boldsymbol{\phi}}(\boldsymbol{x}_i)}{\|f_{\boldsymbol{\phi}}(\boldsymbol{x}_i)\|_2}$ the L2-normalized embedded features, and τ is a temperature parameter.

Now, for each single few-shot task, we introduce our empirical weighted mutual information between the query samples and their latent labels, which integrates two terms: The first is an empirical (Monte-Carlo) estimate of the conditional entropy of labels given the query raw features, denoted $\widehat{\mathcal{H}}(Y_{\mathbb{Q}} | X_{\mathbb{Q}})$, while the second is the empirical label-marginal entropy, $\widehat{\mathcal{H}}(Y_{\mathbb{Q}})$.:

$$\widehat{\mathcal{I}}_\alpha(X_{\mathbb{Q}}; Y_{\mathbb{Q}}) := \alpha \underbrace{\frac{1}{|\mathbb{Q}|} \sum_{i \in \mathbb{Q}} \sum_{k=1}^{K} p_{ik} \log(p_{ik})}_{-\widehat{\mathcal{H}}(Y_{\mathbb{Q}}|X_{\mathbb{Q}}):\ \text{conditional entropy}} \underbrace{- \sum_{k=1}^{K} \widehat{p}_k \log \widehat{p}_k}_{\widehat{\mathcal{H}}(Y_{\mathbb{Q}}):\ \text{marginal entropy}}, \tag{2.2}$$

with α a non-negative hyper-parameter. Notice that setting $\alpha = 1$ recovers the standard mutual information. Setting $\alpha < 1$ allows us to down-weight the conditional entropy term, whose

[3] In order to simplify our notations, we deliberately omit the dependence of posteriors p_{ik} on the network parameters $(\boldsymbol{\phi}, \boldsymbol{\theta})$. Also, p_{ik} takes the form of *softmax* predictions, but we omit the normalization constants.

gradients may dominate the marginal entropy's gradients as the predictions move towards the vertices of the simplex. The role of both terms in (2.2) will be discussed after introducing our overall transductive inference loss in the following, by embedding supervision from the task's support set.

We embed supervision information from support set $\mathbb{S}$ by integrating a standard cross-entropy loss CE with the information measure in Eq. (2.2), which enables us to formulate our Transductive Information Maximization (**TIM**) loss as follows:

$$\boxed{\min_{\boldsymbol{\theta}} \ \lambda \cdot \mathrm{CE} - \widehat{\mathcal{I}}_{\alpha}(X_{\mathbb{Q}}; Y_{\mathbb{Q}})} \quad \text{with} \quad \mathrm{CE} := -\frac{1}{|\mathbb{S}|} \sum_{i \in \mathbb{S}} \sum_{k=1}^{K} y_{ik} \log(p_{ik}), \tag{2.3}$$

where $\{y_{ik}\}$ denotes the k^{th} component of the one-hot encoded label $\boldsymbol{y}_i$ associated to the i-th support sample. Non-negative hyper-parameters α and λ will be fixed to $\alpha = \lambda = 0.1$ in all our experiments. It is worth to discuss in more details the role (importance) of the mutual information terms in (2.3):

- Conditional entropy $\widehat{\mathcal{H}}(Y_{\mathbb{Q}}|X_{\mathbb{Q}})$ aims at minimizing the uncertainty of the posteriors at unlabeled query samples, thereby encouraging the model to output *confident* predictions[4]. This entropy loss is widely used in the context of semi-supervised learning (SSL) (Grandvalet & Bengio, 2005; Miyato *et al.*, 2018; Berthelot *et al.*, 2019), as it models effectively the *cluster* assumption: The classifier's boundaries should not occur at dense regions of the unlabeled features (Grandvalet & Bengio, 2005). Recently, (Dhillon *et al.*, 2020) introduced this term for few-shot learning, showing that entropy fine-tuning on query samples achieves competitive performances. In fact, if we remove the marginal entropy $\widehat{\mathcal{H}}(Y_{\mathbb{Q}})$ in objective (2.3), our TIM objective reduces to the loss in (Dhillon *et al.*, 2020). The conditional entropy $\widehat{\mathcal{H}}(Y_{\mathbb{Q}}|X_{\mathbb{Q}})$ is of paramount importance but its optimization requires special care, as its optima may easily lead to degenerate (non-suitable) solutions on the simplex vertices, mapping all samples to a single class. Such care may consist in using small learning rates and

[4] The global minima of each pointwise entropy in the sum of $\widehat{\mathcal{H}}(Y_{\mathbb{Q}}|X_{\mathbb{Q}})$ are one-hot vectors at the vertices of the simplex.

fine-tuning the whole network (which itself often contains several layers of regularization) as done in (Dhillon *et al.*, 2020), both of which significantly slow down transductive inference.

- The label-marginal entropy regularizer $\widehat{\mathcal{H}}(Y_{\mathbb{Q}})$ encourages the marginal distribution of labels to be uniform, thereby avoiding degenerate solutions obtained when solely minimizing conditional entropy. Hence, it is highly important as it removes the need for implicit regularization, as mentioned in the previous paragraph. In particular, high-accuracy results can be obtained even using higher learning rates and fine-tuning only a fraction of the network parameters (classifier weights $\boldsymbol{\theta}$ instead of the whole network), speeding up substantially transductive runtimes. As it will be observed from our experiments, this term brings substantial improvements in performances (e.g., up to 10% increase in accuracy over entropy fine-tuning on the standard few-shot benchmarks), while facilitating optimization, thereby reducing transductive runtimes by orders of magnitude.

2.2.3 Optimization

At this stage, we consider that the feature extractor has already been trained on base classes (using standard cross-entropy). We now propose two methods for minimizing our objective (2.3) for each test task. The first one is based on standard Gradient Descent (GD). The second is a novel way of optimizing mutual information, and is inspired by the Alternating Direction Method of Multipliers (ADMM). For both methods:

- The pre-trained feature extractor f_ϕ is kept fixed. Only the weights $\boldsymbol{\theta}$ are optimized for each task. Such a choice is discussed in details in subsection 2.3.4. Overall, and interestingly, we found that fine-tuning only classifier weights $\boldsymbol{\theta}$, while fixing feature-extractor parameters ϕ, yielded the best performances for our mutual-information loss.
- For each task, weights $\boldsymbol{\theta}$ are initialized as the class prototypes of the support set:

$$\boldsymbol{\theta}_k^{(0)} = \frac{\sum_{i \in \mathbb{S}} y_{ik} z_i}{\sum_{i \in \mathbb{S}} y_{ik}}$$

Gradient descent (TIM-GD): A straightforward way to minimize our loss in Eq. (2.3) is to perform gradient descent over $\boldsymbol{\theta}$, which we update using all the samples of the few-shot task (both support and query) at once (i.e., no mini-batch sampling). This gradient approach yields our overall best results, while being one order of magnitude faster than the transductive entropy-based fine-tuning in (Dhillon *et al.*, 2020). As will be shown later in our experiments, the method in (Dhillon *et al.*, 2020) needs to fine-tune the whole network (i.e., to update both ϕ and $\boldsymbol{\theta}$), which provides implicit regularization, avoiding the degenerate solutions of entropy minimization. However, TIM-GD (with $\boldsymbol{\theta}$-updates only) still remains two orders of magnitude slower than inductive closed-form solutions (Snell *et al.*, 2017). In the following, we present a more efficient solver for our problem.

Alternating direction method (TIM-ADM): We derive an Alternating Direction Method (ADM) for minimizing our objective in (2.3). Such scheme yields substantial speedups in transductive learning (one order of magnitude), while maintaining the high levels of accuracy of TIM-GD. To do so, we introduce auxiliary variables representing latent assignments of query samples, and minimize a mixed-variable objective by alternating two sub-steps, one optimizing w.r.t classifier's weights $\boldsymbol{\theta}$, and the other w.r.t the auxiliary variables $\boldsymbol{q}$.

Proposition 2.2.0.1. *The objective in (2.3) can be approximately minimized via the following constrained formulation of the problem:*

$$\min_{\boldsymbol{\theta},\boldsymbol{q}} \underbrace{-\frac{\lambda}{|\mathbb{S}|}\sum_{i\in\mathbb{S}}\sum_{k=1}^{K} y_{ik}\log(p_{ik})}_{\text{CE}} + \underbrace{\sum_{k=1}^{K}\widehat{q}_k\log\widehat{q}_k}_{\sim\widehat{\mathcal{H}}(Y_\mathbb{Q})} \underbrace{-\frac{\alpha}{|\mathbb{Q}|}\sum_{i\in\mathbb{Q}}\sum_{k=1}^{K} q_{ik}\log(p_{ik})}_{\sim\widehat{\mathcal{H}}(Y_\mathbb{Q}|X_\mathbb{Q})} + \underbrace{\frac{1}{|\mathbb{Q}|}\sum_{i\in\mathbb{Q}}\sum_{k=1}^{K} q_{ik}\log\frac{q_{ik}}{p_{ik}}}_{\textit{Penalty}\equiv\mathcal{D}_{\mathrm{KL}}(\mathbf{q}\|\mathbf{p})}$$

$$s.t \quad \sum_{k=1}^{K} q_{ik} = 1, \quad q_{ik} \geq 0, \quad i \in \mathbb{Q}, \quad k \in \{1,\dots,K\}, \tag{2.4}$$

where $\boldsymbol{q} = [q_{ik}] \in \mathbb{R}^{|\mathbb{Q}|\times K}$ *are auxiliary variables,* $\boldsymbol{p} = [p_{ik}] \in \mathbb{R}^{|\mathbb{Q}|\times K}$ *and* $\widehat{q}_k = \frac{1}{|\mathbb{Q}|}\sum_{i\in\mathbb{Q}} q_{ik}$*.*

Proof. It is straightforward to notice that, when equality constraints $q_{ik} = p_{ik}$ are satisfied, the last term in objective (2.4), which can be viewed as a soft penalty for enforcing those equality constraints, vanishes. Objectives (2.3) and (2.4) then become equivalent. □

Splitting the problem into sub-problems on $\boldsymbol{\theta}$ and $\mathbf{q}$ as in Eq. (2.4) is closely related to the general principle of ADMM (Alternating Direction Method of Multipliers) (Boyd, Parikh, Chu, Peleato & Eckstein, 2011), except that the KL divergence is not a typical penalty for imposing the equality constraints[5]. The main idea is to **decompose the original problem into two easier sub-problems**, one over $\boldsymbol{\theta}$ and the other over $\boldsymbol{q}$, which can be alternately solved, each in closed-form. Interestingly, this KL penalty is important as it completely removes the need for dual iterations for the simplex constraints in (2.4), yielding closed-form solutions:

Proposition 2.2.0.2. *ADM formulation in Proposition 2.2.0.1 can be approximately solved by alternating the following closed-form updates w.r.t auxiliary variables* $\mathbf{q}$ *and classifier weights* $\boldsymbol{\theta}$ *(t is the iteration index):*

$$q_{ik}^{(t+1)} \propto \frac{\left(p_{ik}^{(t)}\right)^{1+\alpha}}{\left(\sum_{i \in \mathbb{Q}} \left(p_{ik}^{(t)}\right)^{1+\alpha}\right)^{1/2}} \tag{2.5}$$

$$\boldsymbol{\theta}_k^{(t+1)} \leftarrow \frac{\frac{\lambda}{1+\alpha} \sum_{i \in \mathbb{S}} \left(y_{ik}\, z_i + p_{ik}^{(t)} (\boldsymbol{\theta}_k^{(t)} - z_i)\right) + \frac{|\mathbb{S}|}{|\mathbb{Q}|} \sum_{i \in \mathbb{Q}} \left(q_{ik}^{(t+1)} z_i + p_{ik}^{(t)} (\boldsymbol{\theta}_k^{(t)} - z_i)\right)}{\frac{\lambda}{1+\alpha} \sum_{i \in \mathbb{S}} y_{ik} + \frac{|\mathbb{S}|}{|\mathbb{Q}|} \sum_{i \in \mathbb{Q}} q_{ik}^{(t+1)}} \tag{2.6}$$

Proof. A detailed proof is deferred to the supplementary material. Here, we summarize the main technical ingredients of the approximation. Keeping the auxiliary variables $\mathbf{q}$ fixed, we optimize a convex approximation of Eq. (2.4) w.r.t $\boldsymbol{\theta}$. With $\boldsymbol{\theta}$ fixed, the objective is strictly convex w.r.t the auxiliary variables $\mathbf{q}$ whose updates come from a closed-form solution of the KKT (Karush–Kuhn–Tucker) conditions. Interestingly, the negative entropy of auxiliary variables, which appears in the penalty term, handles implicitly the simplex constraints, which removes the need for dual iterations to solve the KKT conditions. □

[5] Typically, ADMM methods use multiplier-based quadratic penalties for enforcing the equality constraint.

2.3 Experiments

Hyperparameters: To keep our experiments as simple as possible, our hyperparameters are kept fixed across all the experiments and methods (TIM-GD and TIM-ADM). The conditional entropy weight α and the cross-entropy weights λ in Objective (2.3) are both set to 0.1. The temperature parameter τ in the classifier is set to 15. In our TIM-GD method, we use the ADAM optimizer with the recommended parameters (Kingma & Ba, 2014), and run 1000 iterations for each task. For TIM-ADM, we run 150 iterations.

Base-training procedure: The feature extractors are trained following the same simple base-training procedure as in (Ziko *et al.*, 2020) and using standard networks (RN-18 and WRN), for all the experiments. Specifically, they are trained using the standard cross-entropy loss on the base classes, with label smoothing, which prevents overfitting and can help obtain more generalizable features. The label-smoothing parameter is set to 0.1. We emphasize that base training does not involve any meta-learning or episodic training strategy. The models are trained for 90 epochs, with the learning rate initialized to 0.1, and divided by 10 at epochs 45 and 66. Batch size is set to 256 for RN-18, and to 128 for WRN. During training, all the images are resized to 84×84, and we used the same data augmentation procedure as in (Ziko *et al.*, 2020), which includes random cropping, color jitter and random horizontal flipping.

Datasets: We resort to 3 few-shot learning datasets to benchmark the proposed models. As standard few-shot benchmarks, we use the ***mini*-Imagenet** (Vinyals *et al.*, 2016b) dataset, with 100 classes split as in (Ravi & Larochelle, 2016), the **Caltech-UCSD Birds 200** (Welinder *et al.*, 2010) (CUB) dataset, with 200 classes, split following (Chen *et al.*, 2019), and finally the larger ***tiered*-Imagenet** dataset, with 608 classes split as in (Ren *et al.*, 2018).

2.3.1 Comparison to state-of-the-art

We first evaluate our methods TIM-GD and TIM-ADM on the widely adopted *mini*-ImageNet, *tiered*-ImageNet and *CUB* benchmark datasets, in the most common 1-shot 5-way and 5-shot 5-way scenarios, with 15 query shots for each class. Results are reported in Table 2.1, and are averaged over 10,000 episodes, following (Wang *et al.*, 2019b). We can observe that both TIM-GD and TIM-ADM yield state-of-the-art performances, consistently across all standard datasets, scenarios and backbones, improving over both transductive and inductive methods, by significant margins.

Table 2.1 Comparison to the state-of-the-art methods on *mini*-ImageNet, *tiered*-Imagenet and CUB. The methods are sub-grouped into transductive and inductive methods, as well as by backbone architecture. Our results (gray-shaded) are averaged over 10,000 episodes. "-" signifies the result is unavailable

			***mini*-ImageNet**		***tiered*-ImageNet**		**CUB**	
Method	Transd.	Backbone	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
MAML (Finn *et al.*, 2017)	✗	RN-18	49.6	65.7	-	-	68.4	83.5
RelatNet (Sung *et al.*, 2018)		RN-18	52.5	69.8	-	-	68.6	84.0
MatchNet (Vinyals *et al.*, 2016b)		RN-18	52.9	68.9	-	-	73.5	84.5
ProtoNet (Snell *et al.*, 2017)		RN-18	54.2	73.4	-	-	73.0	86.6
MTL (Sun, Liu, Chua & Schiele, 2019)		RN-12	61.2	75.5	-	-	-	-
Neg-cosine (Liu *et al.*, 2020a)		RN-18	62.3	80.9	-	-	72.7	89.4
MetaOpt (Lee *et al.*, 2019b)		RN-12	62.6	78.6	66.0	81.6	-	-
SimpleShot (Wang *et al.*, 2019b)		RN-18	62.9	80.0	68.9	84.6	68.9	84.0
Distill (Tian *et al.*, 2020a)		RN-12	64.8	82.1	71.5	86.0	-	-
RelatNet + T (Hou *et al.*, 2019)	✓	RN-12	52.4	65.4	-	-	-	-
ProtoNet + T (Hou *et al.*, 2019)		RN-12	55.2	71.1	-	-	-	-
MatchNet+T (Hou *et al.*, 2019)		RN-12	56.3	69.8	-	-	-	-
TPN (Liu *et al.*, 2019c)		RN-12	59.5	75.7	-	-	-	-
TEAM (Qiao *et al.*, 2019)		RN-18	60.1	75.9	-	-	-	-
Ent-min (Dhillon *et al.*, 2020)		RN-12	62.4	74.5	68.4	83.4	-	-
CAN+T (Hou *et al.*, 2019)		RN-12	67.2	80.6	73.2	84.9	-	-
LaplacianShot (Ziko *et al.*, 2020)		RN-18	72.1	82.3	79.0	86.4	81.0	88.7
TIM-ADM		RN-18	73.6	**85.0**	**80.0**	**88.5**	81.9	90.7
TIM-GD		RN-18	**73.9**	**85.0**	79.9	**88.5**	**82.2**	**90.8**
LEO (Rusu *et al.*, 2019)	✗	WRN	61.8	77.6	66.3	81.4	-	-
SimpleShot (Wang *et al.*, 2019b)		WRN	63.5	80.3	69.8	85.3	-	-
MatchNet (Vinyals *et al.*, 2016b)		WRN	64.0	76.3	-	-	-	-
CC+rot+unlabeled (Gidaris *et al.*, 2019)		WRN	64.0	80.7	70.5	85.0	-	-
FEAT (Ye *et al.*, 2020a)		WRN	65.1	81.1	70.4	84.4	-	-
AWGIM (Guo & Cheung, 2020)	✓	WRN	63.1	78.4	67.7	82.8	-	-
Ent-min (Dhillon *et al.*, 2020)		WRN	65.7	78.4	73.3	85.5	-	-
SIB (Hu *et al.*, 2020)		WRN	70.0	79.2	-	-	-	-
BD-CSPN (Liu *et al.*, 2019a)		WRN	70.3	81.9	78.7	86.92	-	-
LaplacianShot (Ziko *et al.*, 2020)		WRN	74.9	84.1	80.2	87.6	-	-
TIM-ADM		WRN	77.5	87.2	82.0	89.7	-	-
TIM-GD		WRN	**77.8**	**87.4**	**82.1**	**89.8**	-	-

Table 2.2 The results for the domain-shift setting *mini*-Imagenet → CUB, in the 5-shot setting. The results obtained by our models (gray-shaded) are averaged over 10,000 episodes

		***mini*-ImageNet → CUB**
Methods	Backbone	5-shot
MatchNet (Vinyals *et al.*, 2016b)	RN-18	53.1
MAML (Finn *et al.*, 2017)	RN-18	51.3
ProtoNet (Snell *et al.*, 2017)	RN-18	62.0
RelatNet (Sung *et al.*, 2018)	RN-18	57.7
SimpleShot (Wang *et al.*, 2019b)	RN-18	64.0
GNN (Tseng, Lee, Huang & Yang, 2020)	RN-10	66.9
Neg-Cosine (Liu *et al.*, 2020a)	RN-18	67.0
Baseline (Chen *et al.*, 2019)	RN-18	65.6
LaplacianShot (Ziko *et al.*, 2020)	RN-18	66.3
TIM-ADM	RN-18	70.3
TIM-GD	RN-18	**71.0**

2.3.2 Impact of domain-shift

Chen et al. (Chen *et al.*, 2019) recently showed that the performance of most meta-learning methods may drop drastically when a domain-shift exists between the base training data and test data. Surprisingly, the simplest discriminative baseline exhibited the best performance in this case. Therefore, we evaluate our methods in this challenging scenario. To this end, we simulate a domain shift by training the feature encoder on *mini*-Imagenet while evaluating the methods on *CUB*, similarly to the setting introduced in (Chen *et al.*, 2019). TIM-GD and TIM-ADM beat previous methods by significant margins in the domain-shift scenario, consistently with our results in the standard few-shot benchmarks, thereby demonstrating an increased potential of applicability to real-world situations.

2.3.3 Pushing the meta-testing stage

Most few-shot papers only evaluate their method in the usual 5-way scenario. Nevertheless, (Chen *et al.*, 2019) showed that meta-learning methods could be beaten by their discriminative baseline when more ways were introduced in each task. Therefore, we also provide results of our method in the more challenging 10-way and 20-way scenarios on *mini*-ImageNet. These results, which are presented in Table 2.3, show that TIM-GD outperforms other methods by significant margins, in both settings.

Table 2.3 Results for increasing the number of classes on *mini*-ImageNet. The results obtained by our models (gray-shaded) are averaged over 10,000 episodes

		10-way		20-way	
Methods	Backbone	1-shot	5-shot	1-shot	5-shot
MatchNet (Vinyals *et al.*, 2016b)	RN-18	-	52.3	-	36.8
ProtoNet (Snell *et al.*, 2017)	RN-18	-	59.2	-	45.0
RelatNet (Sung *et al.*, 2018)	RN-18	-	53.9	-	39.2
SimpleShot (Wang *et al.*, 2019b)	RN-18	45.1	68.1	32.4	55.4
Baseline (Chen *et al.*, 2019)	RN-18	-	55.0	-	42.0
Baseline++ (Chen *et al.*, 2019)	RN-18	-	63.4	-	50.9
TIM-ADM	RN-18	56.0	**72.9**	**39.5**	58.8
TIM-GD	RN-18	**56.1**	72.8	39.3	**59.5**

2.3.4 Ablation study

Influence of each term: We now assess the impact of each term[6] in our loss in Eq. (2.3) on the final performance of our methods. The results are reported in Table 2.4. We observe that integrating the three terms in our loss consistently outperforms any other configuration. Interestingly, removing the label-marginal entropy, $\widehat{\mathcal{H}}(Y_{\mathbb{Q}})$, reduces significantly the performances in both TIM-GD and TIM-ADM, particularly when only classifier weights $\boldsymbol{\theta}$ are updated and feature extractor $\boldsymbol{\phi}$ is fixed. Such a behavior could be explained by the following fact: the conditional entropy term, $\widehat{\mathcal{H}}(Y_{\mathbb{Q}}|X_{\mathbb{Q}})$, may yield degenerate solutions (assigning all query samples to a single class) on numerous tasks, when used alone. This emphasizes the importance of the label-marginal entropy term $\widehat{\mathcal{H}}(Y_{\mathbb{Q}})$ in our loss (2.3), which acts as a powerful regularizer to prevent such trivial solutions.

Fine-tuning the whole network vs only the classifier weights: While our TIM-GD and TIM-ADM optimize w.r.t $\boldsymbol{\theta}$ and keep base-trained encoder $f_{\boldsymbol{\phi}}$ fixed at inference, the authors of (Dhillon *et al.*, 2020) fine-tuned the whole network $\{\boldsymbol{\theta}, \boldsymbol{\phi}\}$ when performing their transductive

[6] The $\boldsymbol{\theta}$ and $\mathbf{q}$ updates of TIM-ADM associated to each configuration can be found in the supplementary material.

Table 2.4 Ablation study on the effect of each term in our loss in Eq. (2.3), when only the classifier weights are fine-tuned, i.e., updating only $\boldsymbol{\theta}$, and when the whole network is fine-tuned, i.e., updating $\{\boldsymbol{\phi}, \boldsymbol{\theta}\}$. The results are reported for RN-18 as backbone. The same term indexing as in Eq. (2.3) is used here: $\widehat{\mathcal{H}}(Y_{\mathbb{Q}})$: Marginal entropy, $\widehat{\mathcal{H}}(Y_{\mathbb{Q}}|X_{\mathbb{Q}})$: Conditional entropy, CE: Cross-entropy

			***mini*-ImageNet**		***tiered*-ImageNet**		**CUB**	
Method	Param.	Loss	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
TIM-ADM	$\{\boldsymbol{\theta}\}$	CE	60.0	79.6	68.0	84.6	68.6	86.4
		$\text{CE} + \widehat{\mathcal{H}}(Y_{\mathbb{Q}}\|X_{\mathbb{Q}})$	36.0	77.0	48.1	82.5	48.5	86.5
		$\text{CE} - \widehat{\mathcal{H}}(Y_{\mathbb{Q}})$	66.7	82.0	74.0	86.5	74.2	88.3
		$\text{CE} - \widehat{\mathcal{H}}(Y_{\mathbb{Q}}) + \widehat{\mathcal{H}}(Y_{\mathbb{Q}}\|X_{\mathbb{Q}})$	**73.6**	**85.0**	**80.0**	**88.5**	**81.9**	**90.7**
TIM-GD	$\{\boldsymbol{\theta}\}$	CE	60.7	79.4	68.4	84.3	69.6	86.3
		$\text{CE} + \widehat{\mathcal{H}}(Y_{\mathbb{Q}}\|X_{\mathbb{Q}})$	35.3	79.2	45.9	80.6	46.1	85.9
		$\text{CE} - \widehat{\mathcal{H}}(Y_{\mathbb{Q}})$	66.1	81.3	73.4	86.0	73.9	88.0
		$\text{CE} - \widehat{\mathcal{H}}(Y_{\mathbb{Q}}) + \widehat{\mathcal{H}}(Y_{\mathbb{Q}}\|X_{\mathbb{Q}})$	**73.9**	**85.0**	**79.9**	**88.5**	**82.2**	**90.8**
TIM-GD	$\{\boldsymbol{\phi}, \boldsymbol{\theta}\}$	CE	60.8	81.6	65.7	83.5	68.7	87.7
		$\text{CE} + \widehat{\mathcal{H}}(Y_{\mathbb{Q}}\|X_{\mathbb{Q}})$	62.7	81.9	66.9	82.8	72.6	89.0
		$\text{CE} - \widehat{\mathcal{H}}(Y_{\mathbb{Q}})$	62.3	82.7	68.3	85.4	70.7	88.8
		$\text{CE} - \widehat{\mathcal{H}}(Y_{\mathbb{Q}}) + \widehat{\mathcal{H}}(Y_{\mathbb{Q}}\|X_{\mathbb{Q}})$	**67.2**	**84.7**	**73.0**	**86.8**	**76.7**	**90.5**

entropy minimization. To assess both approaches, we add to Table 2.4 a variant of TIM-GD, in which we fine-tune the whole network $\{\boldsymbol{\theta}, \boldsymbol{\phi}\}$, by using the same optimization procedure as in (Dhillon *et al.*, 2020). We found that, besides being much slower, fine-tuning the whole network for our objective in Eq. 2.3 degrades the performances, as also conveyed by the convergence plots in Figure 2.1. Interestingly, when fine-tuning the whole network $\{\boldsymbol{\theta}, \boldsymbol{\phi}\}$, the absence of $\widehat{\mathcal{H}}(Y_{\mathbb{Q}})$ in the entropy-based loss $\text{CE} + \widehat{\mathcal{H}}(Y_{\mathbb{Q}}|X_{\mathbb{Q}})$ does not cause the same drastic drop in performance as observed earlier when optimizing with respect to $\boldsymbol{\theta}$ only. We hypothesize that the network's intrinsic regularization (such as batch normalizations) and the use of small learning rates, as prescribed by (Dhillon *et al.*, 2020), help the optimization process, preventing the predictions from approaching the vertices of the simplex, where entropy's gradients diverge.

2.3.5 Inference runtimes

Transductive methods are generally slower at inference than their inductive counterparts, with run-times that are, typically, several orders of magnitude larger. In Table 2.5, we measure the

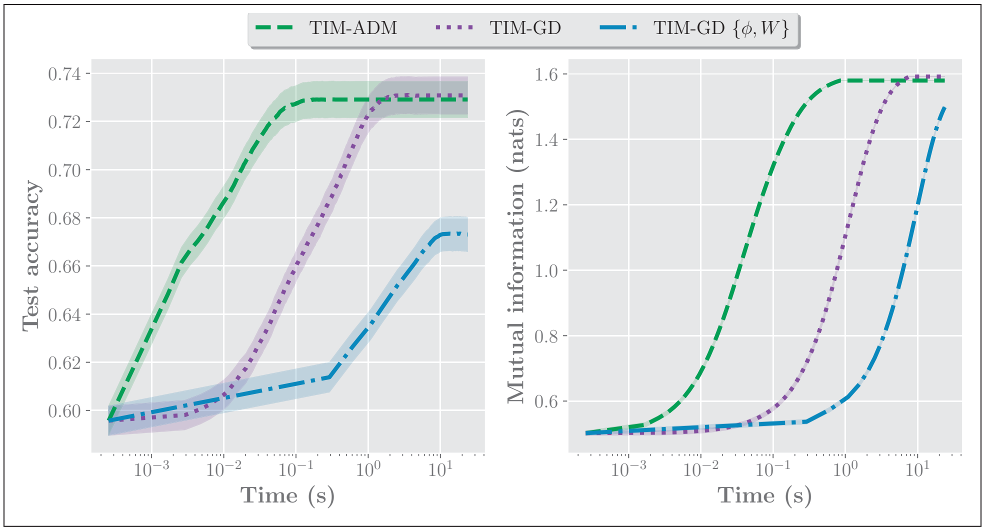

Figure 2.1 Convergence plots for our methods on *mini*-ImageNet with a RN-18. Solid lines are averages, while shadows are 95% confidence intervals. Time is on a logarithmic scale. **Left:** Evolution of the test accuracy during transductive inference. **Right:** Evolution of the mutual information between query features and predictions $\widehat{\mathcal{I}}(X_{\mathbb{Q}}; Y_{\mathbb{Q}})$, computed as in Eq. (2.2), with $\alpha = 1$

Table 2.5 Inference runtime per few-shot task for a 5-shot 5-way task on mini-ImageNet with a WRN backbone

Method	Parameters	Transductive	Inference/task (s)
SimpleShot (Wang *et al.*, 2019b)	$\{\boldsymbol{\theta}\}$	✗	9.0×10^{-3}
TIM-ADM	$\{\boldsymbol{\theta}\}$	✓	1.2×10^{-1}
TIM-GD	$\{\boldsymbol{\theta}\}$		$2.2 \times 10^{+0}$
Ent-min (Dhillon *et al.*, 2020)	$\{\boldsymbol{\phi}, \boldsymbol{\theta}\}$		$2.1 \times 10^{+1}$

average adaptation time per few-shot task, defined as the time required by each method to build the final classifier, for a 5-shot 5-way task on *mini*-ImageNet using the WRN network. Table 2.5 conveys that our ADM optimization gains one order of magnitude in run-time over our gradient-based method, and more than two orders of magnitude in comparison to (Dhillon *et al.*, 2020), which fine-tunes the whole network. Note that TIM-ADM still remains slower than the inductive baseline. Our methods were run on the same GTX 1080 Ti GPU, while the run-time of (Dhillon *et al.*, 2020) is directly reported from the paper.

2.4 Conclusion and future work

Our TIM inference establishes new state-of-the-art results on the standard few-shot benchmarks, as well as in more challenging scenarios, with larger numbers of classes and domain shifts. We used feature extractors based on a simple base-class training with the standard cross-entropy loss, without resorting to the complex meta-training schemes that are often used and advocated in the recent few-shot literature. TIM is modular: it could be plugged on top of any feature extractor and base training, regardless of how the training was conducted. Therefore, while we do not claim that the very challenging few-shot problem is solved, we believe that our model-agnostic TIM inference should be used as a strong baseline for future few-shot learning research. In future work, we target on giving a more theoretical ground for our proposed mutual-information objective, and on exploring further generalizations of the objective, e.g., via embedding domain-knowledge priors. Specifically, one of our theoretical goals will be to connect TIM's objective to the classifier's empirical risk on the query set, showing that the former could be viewed as a surrogate for the latter.

2.5 Acknowledgements

This research was supported by the National Science and Engineering Research Council of Canada (NSERC), via its Discovery Grant program.

CHAPTER 3

REALISTIC EVALUATION OF TRANSDUCTIVE FEW-SHOT LEARNING

Olivier Veilleux[1]∗ , Malik Boudiaf[1]∗ , Pablo Piantanida[2] , Ismail Ben Ayed[1]

[1] ÉTS Montréal, QC, Canada,
[2] Laboratoire des Signaux et Systèmes (L2S),
CentraleSupelec-CNRS-Université Paris-Saclay, France

[*] Equal contributions

Poster at Neural Information Processing Systems (NeurIPS) 2021, Vancouver.

Abstract

Transductive inference is widely used in few-shot learning, as it leverages the statistics of the unlabeled query set of a few-shot task, typically yielding substantially better performances than its inductive counterpart. The current few-shot benchmarks use perfectly class-balanced tasks at inference. We argue that such an artificial regularity is unrealistic, as it assumes that the marginal label probability of the testing samples is known and fixed to the uniform distribution. In fact, in realistic scenarios, the unlabeled query sets come with arbitrary and unknown label marginals. We introduce and study the effect of arbitrary class distributions within the query sets of few-shot tasks at inference, removing the class-balance artifact. Specifically, we model the marginal probabilities of the classes as Dirichlet-distributed random variables, which yields a principled and realistic sampling within the simplex. This leverages the current few-shot benchmarks, building testing tasks with arbitrary class distributions. We evaluate experimentally state-of-the-art transductive methods over 3 widely used data sets, and observe, surprisingly, substantial performance drops, even below inductive methods in some cases. Furthermore, we propose a generalization of the mutual-information loss, based on α-divergences, which can handle effectively class-distribution variations. Empirically, we show that our transductive α-divergence optimization outperforms state-of-the-art methods across several data sets, models and few-shot settings. Our code is publicly available at https://github.com/oveilleux/Realistic_Transductive_Few_Shot.

3.1 Introduction

Deep learning models are widely dominating the field. However, their outstanding performances are often built upon training on large-scale labeled data sets, and the models are seriously challenged when dealing with novel classes that were not seen during training. Few-shot learning (Fei-Fei *et al.*, 2006; Miller, Matsakis & Viola, 2000a; Vinyals *et al.*, 2016b) tackles this challenge, and has recently triggered substantial interest within the community. In standard few-shot settings, a model is initially trained on large-scale data containing labeled examples from a set of *base* classes. Then, supervision for a new set of classes, which are different from those seen in the base training, is restricted to just one or a few labeled samples per class. Model generalization is evaluated over few-shot *tasks*. Each task includes a *query* set containing unlabeled samples for evaluation, and is supervised by a *support* set containing a few labeled samples per new class.

The recent few-shot classification literature is abundant and widely dominated by convoluted meta-learning and episodic-training strategies. To simulate generalization challenges at test times, such strategies build sequences of artificially balanced few-shot tasks (or episodes) during base training, each containing both query and support samples. Widely adopted methods within this paradigm include: Prototypical networks (Snell *et al.*, 2017), which optimizes the log-posteriors of the query points within each base-training episode; Matching networks (Vinyals *et al.*, 2016b), which expresses the predictions of query points as linear functions of the support labels, while deploying episodic training and memory architectures; MAML (Model-Agnostic Meta-Learning) (Finn *et al.*, 2017), which encourages a model to be “easy” to fine-tune; and the meta-learner in (Ravi & Larochelle, 2016), which prescribes optimization as a model for few-shot learning. These popular methods have recently triggered a large body of few-shot learning literature, for instance, (Sung *et al.*, 2018; Oreshkin *et al.*, 2018; Mishra *et al.*, 2018; Rusu *et al.*, 2019; Liu *et al.*, 2019c; Hou *et al.*, 2019; Ye, Hu, Zhan & Sha, 2020b), to list a few.

Recently, a large body of works investigated *transductive* inference for few-shot tasks, e.g., (Liu *et al.*, 2019c; Qiao *et al.*, 2019; Hou *et al.*, 2019; Dhillon *et al.*, 2020; Hu *et al.*, 2020; Wang, Xu, Liu, Zhang & Fu, 2020b; Guo & Cheung, 2020; Yang *et al.*, 2020c; Ziko *et al.*, 2020; Liu *et al.*, 2019b; Liu, Schiele & Sun, 2020e; Boudiaf *et al.*, 2020a), among many others, showing substantial improvements in performances over *inductive* inference[7]. Also, as discussed in (Bronskill, Gordon, Requeima, Nowozin & Turner, 2020), most meta-learning approches rely critically on transductive batch normalization (TBN) to achieve competitive performances, for instance, the methods in (Finn *et al.*, 2017; Gordon, Bronskill, Bauer, Nowozin & Turner, 2019; Zhen *et al.*, 2020), among others. Adopted initially in the widely used MAML (Finn *et al.*, 2017), TBN performs normalization using the statistics of the query set of a given few-shot task, and yields significant increases in performances (Bronskill *et al.*, 2020). Therefore, due to the popularity of MAML, several meta-learning techniques have used TBN. The transductive setting is appealing for few-shot learning, and the outstanding performances observed recently resonate well with a well-known fact in classical transductive inference (Vapnik, 1999; Joachims, 1999; Dengyong *et al.*, 2004): On small labeled data sets, transductive inference outperforms its inductive counterpart. In few-shot learning, transductive inference has access to exactly the same training and testing data as its inductive counterpart[8]. The difference is that it classifies all the unlabeled query samples of each single few-shot task jointly, rather than one sample at a time.

The current few-shot benchmarks use perfectly class-balanced tasks at inference: For each task used at testing, all the classes have exactly the same number of samples, i.e., the marginal probability of the classes is assumed to be known and fixed to the uniform distribution across all tasks. This may not reflect realistic scenarios, in which testing tasks might come with arbitrary class proportions. For instance, the unlabeled query set of a task could be highly imbalanced. In fact, using perfectly balanced query sets for benchmarking the models assumes exact knowledge of the marginal distributions of the true labels of the testing points, but such labels are unknown. This is, undeniably, an unrealistic assumption and an important limitation

[7] The best-performing state-of-the-art few-shot methods in the transductive-inference setting have achieved performances that are up to 10% higher than their inductive counterparts; see (Boudiaf *et al.*, 2020a), for instance.

[8] Each single few-shot task is treated independently of the other tasks in the transductive-inference setting. Hence, the setting does not use additional unlabeled data, unlike semi-supervised few-shot learning (Ren *et al.*, 2018).

of the current few-shot classification benchmarks and datasets. Furthermore, this suggests that the recent progress in performances might be, in part, due to class-balancing priors (or biases) that are encoded in state-of-the-art transductive models. Such priors might be implicit, e.g., through carefully designed episodic-training schemes and specialized architectures, or explicit, e.g., in the design of transductive loss functions and constraints. For instance, the best performing methods in (Boudiaf *et al.*, 2020a; Hu *et al.*, 2021) use explicit label-marginal terms or constraints, which strongly enforce perfect class balance within the query set of each task. In practice, those class-balance priors and assumptions may limit the applicability of the existing few-shot benchmarks and methods. In fact, our experiments show that, over few-shot tasks with random class balance, the performances of state-of-the-art methods may decrease by margins. This motivates re-considering the existing benchmarks and re-thinking the relevance of class-balance biases in state-of-the-art methods.

Contributions

We introduce and study the effect of arbitrary class distributions within the query sets of few-shot tasks at inference. Specifically, we relax the assumption of perfectly balanced query sets and model the marginal probabilities of the classes as Dirichlet-distributed random variables. We devise a principled procedure for sampling simplex vectors from the Dirichlet distribution, which is widely used in Bayesian statistics for modeling categorical events. This leverages the current few-shot benchmarks by generating testing tasks with arbitrary class distributions, thereby reflecting realistic scenarios. We evaluate experimentally state-of-the-art transductive few-shot methods over 3 widely used datasets, and observe that the performances decrease by important margins, albeit at various degrees, when dealing with arbitrary class distributions. In some cases, the performances drop even below the inductive baselines, which are not affected by class-distribution variations (as they do not use the query-set statistics). Furthermore, we propose a generalization of the transductive mutual-information loss, based on α-divergences, which can handle effectively class-distribution variations. Empirically, we show that our transductive α-divergence optimization outperforms state-of-the-art few-shot methods across different data sets, models and few-shot settings.

3.2 Standard few-shot settings

3.2.1 Base training

Assume that we have access to a fully labelled *base* dataset $\mathcal{D}_{base} = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{N_{base}}$, where $\boldsymbol{x}_i \in \mathcal{X}_{base}$ are data points in an input space $\mathcal{X}_{base}$, $\boldsymbol{y}_i \in \{0,1\}^{|\mathcal{Y}_{base}|}$ the one-hot labels, and $\mathcal{Y}_{base}$ the set of base classes. Base training learns a feature extractor $f_{\boldsymbol{\phi}} : \mathcal{X} \to \mathcal{Z}$, with $\boldsymbol{\phi}$ its learnable parameters and $\mathcal{Z}$ a (lower-dimensional) feature space. The vast majority of the literature adopts episodic training at this stage, which consists in formatting $\mathcal{D}_{base}$ as a series of tasks (=episodes) in order to mimic the testing stage, and train a meta-learner to produce, through a differentiable process, predictions for the query set. However, it has been repeatedly demonstrated over the last couple years that a standard supervised training followed by standard transfer learning strategies actually outperforms most meta-learning based approaches (Chen *et al.*, 2019; Wang *et al.*, 2019b; Tian *et al.*, 2020a; Ziko *et al.*, 2020; Boudiaf *et al.*, 2020a). Therefore, we adopt a standard cross-entropy training in this work.

3.2.2 Testing

The model is evaluated on a set of few-shot tasks, each formed with samples from $\mathcal{D}_{test} = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{N_{test}}$, where $\boldsymbol{y}_i \in \{0,1\}^{|\mathcal{Y}_{test}|}$ such that $\mathcal{Y}_{base} \cap \mathcal{Y}_{test} = \emptyset$. Each task is composed of a labelled support set $\mathbb{S} = \{\boldsymbol{x}_i, y_i\}_{i \in I_{\mathbb{S}}}$ and an unlabelled query set $\mathbb{Q} = \{\boldsymbol{x}_i\}_{i \in I_{\mathbb{Q}}}$, both containing instances only from K distinct classes randomly sampled from $\mathcal{Y}_{test}$, with $K < |\mathcal{Y}_{test}|$. Leveraging a feature extractor $f_{\boldsymbol{\phi}}$ pre-trained on the base data, the objective is to learn, for each few-shot task, a classifier $f_{\boldsymbol{\theta}} : \mathcal{Z} \to \Delta_K$, with $\boldsymbol{\theta}$ the learnable parameters and $\Delta_K = \{\boldsymbol{y} \in [0,1]^K \ / \ \sum_k y_k = 1\}$ the $(K-1)$-simplex. To simplify the equations for the rest of the paper, we use the following notations for the posterior predictions of each $i \in I_{\mathbb{S}} \cup I_{\mathbb{Q}}$ and for the class marginals within $\mathbb{Q}$:

$$p_{ik} = f_{\boldsymbol{\theta}}(f_{\boldsymbol{\phi}}(\boldsymbol{x}_i))_k = \mathbb{P}(Y = k | X = \boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\phi}) \ \text{ and } \ \widehat{p}_k = \frac{1}{|I_{\mathbb{Q}}|} \sum_{i \in I_{\mathbb{Q}}} p_{ik} = \mathbb{P}(Y_{\boldsymbol{\theta}} = k; \boldsymbol{\theta}, \boldsymbol{\phi})$$

wThe end goal is to predict the classes of the unlabeled samples in $\mathbb{Q}$ for each few-shot task, independently of the other tasks. A large body of works followed a transductive-prediction setting, e.g., (Liu *et al.*, 2019c; Qiao *et al.*, 2019; Hou *et al.*, 2019; Dhillon *et al.*, 2020; Hu *et al.*, 2020; Wang *et al.*, 2020b; Guo & Cheung, 2020; Yang *et al.*, 2020c; Ziko *et al.*, 2020; Liu *et al.*, 2019b, 2020e; Boudiaf *et al.*, 2020a), among many others. Transductive inference performs a joint prediction for all the unlabeled query samples of each single few-shot task, thereby leveraging the query-set statistics. On the current benchmarks, tranductive inference often outperforms substantially its inductive counterpart (i.e., classifying one sample at a time for a given task). Note that our method is agnostic to the specific choice of classifier $f_{\boldsymbol{\theta}}$, whose parameters are learned at inference. In the experimental evaluations of our method, similarly to (Boudiaf *et al.*, 2020a), we used $p_{ik} \propto \exp(-\frac{\tau}{2}\, \|\boldsymbol{\theta}_k - z_i\|^2)$, with $\boldsymbol{\theta} := (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)$, $z_i = \frac{f_{\boldsymbol{\phi}}(\boldsymbol{x}_i)}{\|f_{\boldsymbol{\phi}}(\boldsymbol{x}_i)\|_2}$, τ is a temperature parameter and base-training parameters $\boldsymbol{\phi}$ are fixed[9].

3.2.3 Perfectly balanced vs imbalanced tasks

In standard K-way few-shot settings, the support and query sets of each task $\mathcal{T}$ are formed using the following procedure: (i) Randomly sample K classes $\mathcal{Y}_{\mathcal{T}} \subset \mathcal{Y}_{test}$; (ii) For each class $k \in \mathcal{Y}_{\mathcal{T}}$, randomly sample $n_k^{\mathbb{S}}$ support examples, such that $n_k^{\mathbb{S}} = |\mathbb{S}|/K$; and (iii) For each class $k \in \mathcal{Y}_{\mathcal{T}}$, randomly sample $n_k^{\mathbb{Q}}$ query examples, such that $n_k^{\mathbb{Q}} = |\mathbb{Q}|/K$. Such a setting is undeniably artificial as we assume $\mathbb{S}$ and $\mathbb{Q}$ have the same perfectly balanced class distribution. Several recent works (Triantafillou *et al.*, 2020; Lee *et al.*, 2019a; Chen, Dai, Li, Gao & Song, 2020a; Ochal, Patacchiola, Storkey, Vazquez & Wang, 2021) studied class imbalance exclusively on the support set $\mathbb{S}$. This makes sense as, in realistic scenarios, some classes might have more labelled samples than others. However, even these works rely on the assumption that query set $\mathbb{Q}$ is perfectly balanced. We argue that such an assumption is not realistic, as one typically has even less control over the class distribution of $\mathbb{Q}$ than it has over that of $\mathbb{S}$. For the labeled support $\mathbb{S}$, the class distribution is at least fully known and standard strategies from imbalanced

[9] $\boldsymbol{\phi}$ is either fixed, e.g., (Boudiaf *et al.*, 2020a), or fine-tuned during inference, e.g., (Dhillon *et al.*, 2020). There is, however, evidence in the literature that freezing $\boldsymbol{\phi}$ yields better performances (Boudiaf *et al.*, 2020a; Chen *et al.*, 2019; Tian *et al.*, 2020a; Wang *et al.*, 2019b), while reducing the inference time.

supervised learning could be applied (Ochal *et al.*, 2021). This does not hold for $\mathbb{Q}$, for which we need to make class predictions at testing time and whose class distribution is unknown. In fact, generating perfectly balanced tasks at test times for benchmarking the models assumes that one has access to the unknown class distributions of the query points, which requires access to their unknown labels. More importantly, artificial balancing of $\mathbb{Q}$ is implicitly or explicitly encoded in several transductive methods, which use the query set statistics to make class predictions, as will be discussed in section 3.4.

3.3 Dirichlet-distributed class marginals for few-shot query sets

Standard few-shot settings assume that p_k, the proportion of the query samples belonging to a class k within a few-shot task, is deterministic (fixed) and known *priori*: $p_k = n_k^Q/|Q| = 1/K$, for all k and all few-shot tasks. We propose to relax this unrealistic assumption, and to use the Dirichlet distribution to model the proportions (or marginal probabilities) of the classes in few-shot query sets as random variables. Dirichlet distributions are widely used in Bayesian statistics to model K-way categorical events[10]. The domain of the Dirichlet distribution is the set of K-dimensional discrete distributions, i.e., the set of vectors in $(K-1)$-simplex $\Delta_K = \{\boldsymbol{p} \in [0,1]^K \mid \sum_k p_k = 1\}$. Let P_k denotes a random variable associated with class probability p_k, and P the random simplex vector given by $P = (P_1, \ldots, P_K)$. We assume that P follows a Dirichlet distribution with parameter vector $\boldsymbol{a} = (a_1, \ldots, a_K) \in \mathbb{R}^K$: $P \sim \mathtt{Dir}(\boldsymbol{a})$. The Dirichlet distribution has the following density function: $f_{\mathtt{Dir}}(\boldsymbol{p}; \boldsymbol{a}) = \frac{1}{B(\boldsymbol{a})} \prod_{k=1}^K p_k^{a_k - 1}$ for $\boldsymbol{p} = (p_1, \ldots, p_K) \in \Delta_K$, with B denoting the multivariate Beta function, which could be expressed with the Gamma function[11]: $B(\boldsymbol{a}) = \frac{\prod_{k=1}^K \Gamma(a_k)}{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}$.

Figure 3.1 illustrates the Dirichlet density for $K = 3$, with a 2-simplex support represented with an equilateral triangle, whose vertices are probability vectors $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$. We show the density for $\boldsymbol{a} = a\mathbb{1}_K$, with $\mathbb{1}_K$ the K-dimensional vector whose all components are

[10] Note that the Dirichlet distribution is the conjugate prior of the categorical and multinomial distributions.

[11] The Gamma function is given by: $\Gamma(a) = \int_0^\infty t^{a-1} \exp(-t) dt$ for $a > 0$. Note that $\Gamma(a) = (a-1)!$ when a is a strictly positive integer.

equal to 1 and concentration parameter a equal to 0.5, 2, 5 and 50. Note that the limiting case $a \rightarrow +\infty$ corresponds to the standard settings with perfectly balanced tasks, where only uniform distribution, i.e., the point in the middle of the simplex, could occur as marginal distribution of the classes.

The following result, well-known in the literature of random variate generation (Devroye, 1986), suggests that one could generate samples from the multivariate Dirichlet distribution via simple and standard univariate Gamma generators.

Theorem 3.3.1. *((Devroye, 1986, p. 594)) Let $N_1, \ldots, N_K$ be K independent Gamma-distributed random variables with parameters a_k: $N_k \sim \mathtt{Gamma}(a_k)$, i.e., the probability density of N_k is univariate Gamma*[12]*, with shape parameter a_k. Let $P_k = \frac{N_k}{\sum_{k=1}^{K} N_k}, k = 1, \ldots, K$. Then, $P = (P_1, \ldots, P_K)$ is Dirichlet distributed: $P \sim \mathtt{Dir}(\boldsymbol{a})$, with $\boldsymbol{a} = (a_1, \ldots, a_K)$.*

A proof based on the Jacobian of random-variable transformations $P_k = \frac{N_k}{\sum_{k=1}^{K} N_k}$, $k = 1, \ldots, K$, could be found in (Devroye, 1986), p. 594. This result prescribes the following simple procedure for sampling random simplex vectors $(p_1, \ldots, p_K)$ from the multivariate Dirichlet distribution with parameters $\boldsymbol{a} = (a_1, \ldots, a_K)$: First, we draw K independent random samples $(n_1, \ldots, n_K)$ from Gamma distributions, with each n_k drawn from univariate density $f_{\mathtt{Gamma}}(n; a_k)$; To do so, one could use standard random generators for the univariate Gamma density; see Chapter 9 in (Devroye, 1986). Then, we set $p_k = \frac{n_k}{\sum_{k=1}^{K} n_k}$. This enables to generate randomly n_k^{Q}, the number of samples of class k within query set Q, as follows: n_k^{Q} is the closest integer to $p_k|Q|$ such that $\sum_k n_k^{Q} = |Q|$.

3.4 On the class-balance bias of the best-performing few-shot methods

As briefly evoked in section 3.2, the strict balancing of the classes in both $\mathbb{S}$ and $\mathbb{Q}$ represents a strong inductive bias, which few-shot methods can either meta-learn during training or leverage at inference. In this section, we explicitly show how such a class-balance prior is encoded in

[12] Univariate Gamma density is given by: $f_{\mathtt{Gamma}}(n; a_k) = \frac{n^{a_k - 1} \exp(-n)}{\Gamma(a_k)}$, $n \in \mathbb{R}$.

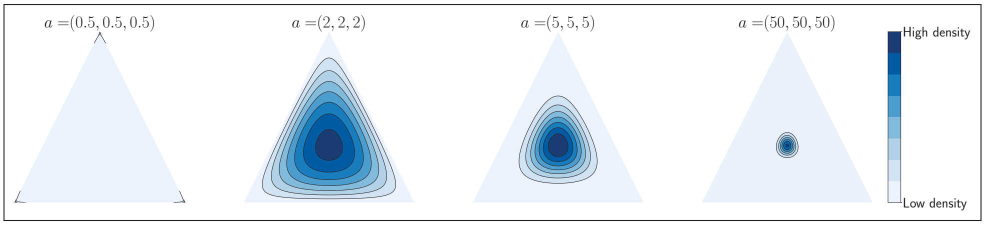

Figure 3.1 Dirichlet density function for $K = 3$, with different choices of parameter vector $\boldsymbol{a}$

the two best-performing transductive methods in the literature (Boudiaf *et al.*, 2020a; Hu *et al.*, 2021), one based on mutual-information maximization (Boudiaf *et al.*, 2020a) and the other on optimal transport (Hu *et al.*, 2021).

3.4.1 Class-balance bias of optimal transport

Recently, the transductive method in (Hu *et al.*, 2021), referred to as PT-MAP, achieved the best performances reported in the literature on several popular benchmarks, to the best of our knowledge. However, the method explicitly embeds a class-balance prior. Formally, the objective is to find, for each few-shot task, an optimal mapping matrix $\boldsymbol{M} \in \mathbb{R}_+^{|\mathbb{Q}|\times K}$, which could be viewed as a joint probability distribution over $X_\mathbb{Q} \times Y_\mathbb{Q}$. At inference, a hard constraint $\boldsymbol{M} \in \{\boldsymbol{M} : \boldsymbol{M}\mathbb{1}_K = \boldsymbol{r}, \mathbb{1}_{|\mathbb{Q}|}\boldsymbol{M} = \boldsymbol{c}\}$ for some $\boldsymbol{r}$ and $\boldsymbol{c}$ is enforced through the use of the Sinkhorn-Knopp algorithm. In other words, the columns and rows of $\boldsymbol{M}$ are constrained to sum to pre-defined vectors $\boldsymbol{r} \in \mathbb{R}^{|\mathbb{Q}|}$ and $\boldsymbol{c} \in \mathbb{R}^K$. Setting $\boldsymbol{c} = \frac{1}{K}\mathbb{1}_K$ as done in (Hu *et al.*, 2021) ensures that $\boldsymbol{M}$ defines a valid joint distribution, *but also crucially encodes the strong prior that all the classes within the query sets are equally likely*. Such a hard constraint is detrimental to the performance in more realistic scenarios where the class distributions of the query sets could be arbitrary, and not necessarily uniform. Unsurprisingly, PT-MAP undergoes a substantial performance drop in the realistic scenario with Dirichlet-distributed class proportions, with a consistent decrease in accuracy between 18 and 20 % on all benchmarks, in the 5-ways case.

3.4.2 Class-balance bias of transductive mutual-information maximization

Let us now have a closer look at the mutual-information maximization in (Boudiaf *et al.*, 2020a). Following the notations introduced in section 3.2, the transductive loss minimized in (Boudiaf *et al.*, 2020a) for a given few-shot task reads:

$$\mathcal{L}_{\text{TIM}} = \text{CE} - \mathcal{I}(X_{\mathbb{Q}}; Y_{\mathbb{Q}}) = \text{CE} \underbrace{- \frac{1}{|I_{\mathbb{Q}}|} \sum_{i \in \mathbb{Q}} \sum_{k=1}^{K} p_{ik} \log(p_{ik})}_{\mathcal{H}(Y_{\mathbb{Q}}|X_{\mathbb{Q}})} + \lambda \underbrace{\sum_{k=1}^{K} \widehat{p}_k \log \widehat{p}_k}_{-\mathcal{H}(Y_{\mathbb{Q}})}, \tag{3.1}$$

where $\mathcal{I}(X_{\mathbb{Q}}; Y_{\mathbb{Q}}) = -\mathcal{H}(Y_{\mathbb{Q}}|X_{\mathbb{Q}}) + \lambda \mathcal{H}(Y_{\mathbb{Q}})$ is a weighted mutual information between the query samples and their unknown labels (the mutual information corresponds to $\lambda = 1$), and $\text{CE} := -\frac{1}{|I_{\mathbb{S}}|} \sum_{i \in \mathbb{S}} \sum_{k=1}^{K} y_{ik} \log(p_{ik})$ is a supervised cross-entropy term defined over the support samples. Let us now focus our attention on the label-marginal entropy term, $\mathcal{H}(Y_{\mathbb{Q}})$. As mentioned in (Boudiaf *et al.*, 2020a), this term is of significant importance as it prevents trivial, single-class solutions stemming from minimizing only conditional entropy $\mathcal{H}(Y_{\mathbb{Q}}|X_{\mathbb{Q}})$. However, we argue that this term also encourages class-balanced solutions. In fact, we can write it as an explicit KL divergence, which penalizes deviation of the label marginals within a query set from the uniform distribution:

$$\mathcal{H}(Y_{\mathbb{Q}}) = -\sum_{k=1}^{K} \widehat{p}_k \log(\widehat{p}_k) = \log(K) - \mathcal{D}_{\text{KL}}(\widehat{\boldsymbol{p}} \| \mathbf{u}_K). \tag{3.2}$$

Therefore, minimizing marginal entropy $\mathcal{H}(Y_{\mathbb{Q}})$ is equivalent to minimizing the KL divergence between the predicted marginal distribution $\widehat{\boldsymbol{p}} = (\widehat{p}_1, \ldots, \widehat{p}_K)$ and uniform distribution $\mathbf{u}_K = \frac{1}{K} \mathbb{1}_K$. This KL penalty could harm the performances whenever the class distribution of the few-shot task is no longer uniform. In line with this analysis, and unsurprisingly, we observe in section 3.6 that the original model in (Boudiaf *et al.*, 2020a) also undergoes a dramatic performance drop, up to 20%. While naively removing this marginal-entropy term leads to even worse performances, we observe that simply down-weighting it, i.e., decreasing λ in Eq.

(3.1), can drastically alleviate the problem, in contrast to the case of optimal transport where the class-balance constraint is enforced in a hard manner.

3.5 Generalizing mutual information

In this section, we propose a non-trivial, but simple generalization of the mutual-information loss in (3.1), based on α-divergences, which can tolerate more effectively class-distribution variations. We identified in section 3.4 a class-balance bias encoded in the marginal Shannon entropy term. Ideally, we would like to extend this Shannon-entropy term in a way that allows for more flexibility: Our purpose is to control how far the predicted label-marginal distribution, $\widehat{\boldsymbol{p}}$, could depart from the uniform distribution without being heavily penalized.

3.5.1 Background

We argue that such flexibility could be controlled through the use of α-divergences (Chernoff *et al.*, 1952; Amari, 2000; Havrda & Charvát, 1967; Tsallis, 1988; Cichocki & Amari, 2010), which generalize the well-known and widely used KL divergence. α-divergences form a whole family of divergences, which encompasses Tsallis and Renyi α-divergences, among others. In this work, we focus on Tsallis's (Chernoff *et al.*, 1952; Tsallis, 1988) formulation of α-divergence. Let us first introduce the generalized logarithm (Cichocki & Amari, 2010): $\log_\alpha(x) = \frac{1}{1-\alpha}\left(x^{1-\alpha} - 1\right)$ Using the latter, Tsallis α-divergence naturally extends KL. For two discrete distributions $\boldsymbol{p} = (p_k)_{k=1}^K$ and $\boldsymbol{q} = (q_k)_{k=1}^K$, we have:

$$\mathcal{D}_\alpha(\boldsymbol{p}\|\boldsymbol{q}) = -\sum_{k=1}^{K} p_k \log_\alpha\left(\frac{q_k}{p_k}\right) = \frac{1}{1-\alpha}\left(1 - \sum_{k=1}^{K} p_k^\alpha q_k^{1-\alpha}\right) \tag{3.3}$$

Note that the Shannon entropy in Eq. (3.2) elegantly generalizes to Tsallis α-entropy:

$$\mathcal{H}_\alpha(\boldsymbol{p}) = \log_\alpha(K) - K^{1-\alpha}\,\mathcal{D}_\alpha(\boldsymbol{p}\|\mathbf{u}_K) = \frac{1}{\alpha-1}\left(1 - \sum_k p_k^\alpha\right) \tag{3.4}$$

The derivation of Eq. (3.4) is provided in the appendix. Also, $\lim_{\alpha \to 1} \log_\alpha(x) = \log(x)$, which implies:

$$\lim_{\alpha \to 1} \mathcal{D}_\alpha(\boldsymbol{p} \| \boldsymbol{q}) = \mathcal{D}_{\text{KL}}(\boldsymbol{p} \| \boldsymbol{q}) \quad \text{and} \quad \lim_{\alpha \to 1} \mathcal{H}_\alpha(\boldsymbol{p}) = \mathcal{H}(\boldsymbol{p}) = -\sum_{k=1}^{K} \widehat{p}_k \log(\widehat{p}_k)$$

Note that α-divergence $\mathcal{D}_\alpha(\boldsymbol{p} \| \boldsymbol{q})$ inherits the nice properties of the KL divergence, including but not limited to convexity with respect to both $\boldsymbol{p}$ and $\boldsymbol{q}$ and strict positivity $\mathcal{D}_\alpha(\boldsymbol{p} \| \boldsymbol{q}) \geq 0$ with equality if $\boldsymbol{p} = \boldsymbol{q}$. Furthermore, beyond its link to the forward KL divergence $\mathcal{D}_{\text{KL}}(\boldsymbol{p} \| \boldsymbol{q})$, α-divergence smoothly connects several well-known divergences, including the reverse KL divergence $\mathcal{D}_{\text{KL}}(\boldsymbol{q} \| \boldsymbol{p})$ ($\alpha \to 0$), the Hellinger ($\alpha = 0.5$) and the Pearson Chi-square ($\alpha = 2$) distances (Cichocki & Amari, 2010).

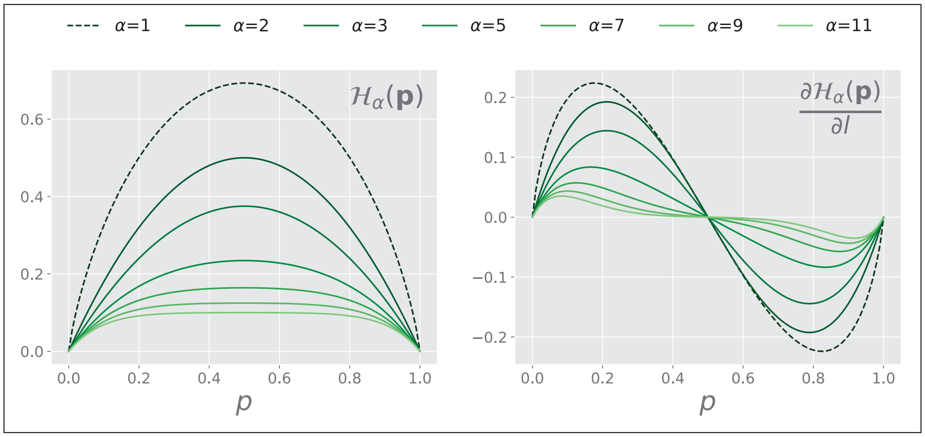

Figure 3.2 (Left) α-entropy as a function of $p = \sigma(l)$. (Right) Gradient of α-entropy w.r.t to the logit $l \in \mathbb{R}$ as a function of $p = \sigma(l)$. Best viewed in color

3.5.2 Analysis of the gradients

As observed from Eq. (3.4), α-entropy is, just like Shannon Entropy, intrinsically biased toward the uniform distribution. Therefore, we still have not properly answered the question: why would α-entropy be better suited to imbalanced situations? We argue the that learning dynamics subtly but crucially differ. To illustrate this point, let us consider a simple toy logistic-regression example. Let $l \in \mathbb{R}$ denotes a logit, and $p = \sigma(l)$ the corresponding probability, where σ stands for the usual sigmoid function. The resulting probability distribution simply reads $\boldsymbol{p} = \{p, 1-p\}$. In Figure 3.2, we plot both the α-entropy $\mathcal{H}_\alpha$ (left) and its gradients $\partial\mathcal{H}_\alpha/\partial l$ (right) as functions of p. The advantage of α-divergence now becomes clearer: as α increases, $\mathcal{H}_\alpha(p)$ accepts more and more deviation from the uniform distribution ($p = 0.5$ on Figure 3.2), while still providing a *barrier* preventing trivial solutions (i.e., $p = 0$ or $p = 1$, which corresponds to all the samples predicted as 0 or 1). Intuitively, such a behavior makes α-entropy with $\alpha > 1$ better suited to class imbalance than Shannon entropy.

3.5.3 Proposed formulation

In light of the previous discussions, we advocate a new α-mutual information loss, a simple but very effective extension of the Shannon mutual information in Eq. (3.1):

$$\mathcal{I}_\alpha(X_\mathbb{Q}; Y_\mathbb{Q}) = \mathcal{H}_\alpha(Y_\mathbb{Q}) - \mathcal{H}_\alpha(Y_\mathbb{Q}|X_\mathbb{Q}) = \frac{1}{\alpha - 1}\left(\frac{1}{|I_\mathbb{Q}|}\sum_{i \in I_\mathbb{Q}}\sum_{k=1}^{K} p_{ik}^{\alpha} - \sum_{k=1}^{K} \widehat{p}_k^{\,\alpha}\right) \tag{3.5}$$

with $\mathcal{H}_\alpha$ the α-entropy as defined in Eq. (3.4). Note that our generalization in Eq. (3.5) has no link to the $\boldsymbol{\alpha}$-mutual information derived in (Arimoto, 1977). Finally, our loss for transductive few-shot inference reads:

$$\mathcal{L}_{\alpha\text{-TIM}} = \text{CE} - \ \mathcal{I}_\alpha(X_\mathbb{Q}; Y_\mathbb{Q}) \tag{3.6}$$

3.6 Experiments

In this section, we thoroughly evaluate the most recent few-shot transductive methods using our imbalanced setting. Except for SIB (Hu *et al.*, 2020) and LR-ICI (Wang *et al.*, 2020b) all the methods have been reproduced in our common framework. All the experiments have been executed on a single GTX 1080 Ti GPU.

3.6.1 Setup

Datasets. We use three standard benchmarks for few-shot classification: *mini*-Imagenet (Vinyals *et al.*, 2016b), *tiered*-Imagenet (Ren *et al.*, 2018) and *Caltech-UCSD Birds 200* (*CUB*) (Welinder *et al.*, 2010). The *mini*-Imagenet benchmark is a subset of the ILSVRC-12 dataset (Deng *et al.*, 2009), composed of 60,000 color images of size 84 x 84 pixels (Vinyals *et al.*, 2016b). It includes 100 classes, each having 600 images. In all experiments, we used the standard split of 64 base-training, 16 validation and 20 test classes (Ravi & Larochelle, 2016; Wang *et al.*, 2019b). The *tiered*-Imagenet benchmark is a larger subset of ILSVRC-12, with 608 classes and 779,165 color images of size 84×84 pixels. We used a standard split of 351 base-training, 97 validation and 160 test classes. The *Caltech-UCSD Birds 200* (*CUB*) benchmark also contains images of size 84×84 pixels, with 200 classes. For CUB, we used a standard split of 100 base-training, 50 validation and 50 test classes, as in (Chen *et al.*, 2019). It is important to note that for all the splits and data-sets, the base-training, validation and test classes are all different.

Task sampling. We evaluate all the methods in the 1-shot 5-way, 5-shot 5-way, 10-shot 5-way and 20-shot 5-way scenarios, with the classes of the query sets randomly distributed following Dirichlet's distribution, as described in section 3.3. Note that the total amount of query samples $|\mathbb{Q}|$ remains fixed to 75. All the methods are evaluated by the average accuracy over 10,000 tasks, following (Wang *et al.*, 2019b). We used different Dirichlet's concentration parameter $\boldsymbol{a}$ for validation and testing. The validation-task generation is based on a random sampling within the simplex (i.e Dirichlet with $\boldsymbol{a} = \mathbb{1}_K$). Testing-task generation uses $\boldsymbol{a} = 2 \cdot \mathbb{1}_K$ to reflect the fact that extremely imbalanced tasks (i.e., only one class is present in the task) are unlikely to happen in practical scenarios; see Figure 3.1 for visualization.

Hyper-parameters. Unless identified as directly linked to a class-balance bias, all the hyper-parameters are kept similar to the ones prescribed in the original papers of the reproduced methods. For instance, the marginal entropy in TIM (Boudiaf *et al.*, 2020a) was identified in section 3.4 as a penalty that encourages class balance. Therefore, the weight controlling the relative importance of this term is tuned. For all methods, hyper-parameter tuning is performed on the validation set of each dataset, using the validation sampling described in the previous paragraph.

Base-training procedure. All non-episodic methods use the same feature extractors, which are trained using the same procedure as in (Boudiaf *et al.*, 2020a; Ziko *et al.*, 2020), via a standard cross-entropy minimization on the base classes with label smoothing. The feature extractors are trained for 90 epochs, using a learning rate initialized to 0.1 and divided by 10 at epochs 45 and 66. We use a batch size of 256 for ResNet-18 and of 128 for WRN28-10. During training, color jitter, random cropping and random horizontal flipping augmentations are applied. For episodic/meta-learning methods, given that each requires a specific training, we use the pre-trained models provided with the GitHub repository of each method.

3.6.2 Main results

The main results are reported in Table 3.1. As baselines, we also report the performances of state-of-the-art inductive methods that do not use the statistics of the query set at adaptation and are, therefore, unaffected by class imbalance. In the 1-shot scenario, all the transductive methods, without exception, undergo a significant drop in performances as compared to the balanced setting. Even though the best-performing transductive methods still outperforms the inductive ones, we observe that more than half of the transductive methods evaluated perform overall worse than inductive baselines in our realistic setting. Such a surprising finding highlights that the standard benchmarks, initially developed for the inductive setting, are not well suited to evaluate transductive methods. In particular, when evaluated with our protocol, the current state-of-the-art holder PT-MAP averages more than 18% performance drop across datasets and backbones, Entropy-Min around 7%, and TIM around 4%. Our proposed α-TIM method outperforms transductive methods across almost all task formats, datasets and backbones, and is the only method that consistently inductive baselines in fair setting. While stronger inductive baselines have been proposed in the literature (Zhang, Cai, Lin & Shen, 2020a), we show in the

Table 3.1 Comparisons of state-of-the-art methods in our realistic setting on *mini*-ImageNet, *tiered*-ImageNet and CUB. Query sets are sampled following a Dirichlet distribution with $\boldsymbol{a} = 2 \cdot \mathbb{1}_K$. Accuracy is averaged over 10,000 tasks. A red arrow (↓) indicates a performance drop between the artificially-balanced setting and our testing procedure, and a blue arrow (↑) an improvement. Arrows are not displayed for the inductive methods as, for these, there is no significant change in performance between both settings (expected). '–' signifies the result was computationally intractable to obtain

			***mini*-ImageNet**			
	Method	**Network**	**1-shot**	**5-shot**	**10-shot**	**20-shot**
Induct.	Protonet (NeurIPS'17 (Snell et al., 2017))	RN-18	53.4	74.2	79.2	82.4
	Baseline (ICLR'19 (Chen et al., 2019))		56.0	78.9	83.2	85.9
	Baseline++ (ICLR'19 (Chen et al., 2019))		60.4	79.7	83.8	86.3
	Simpleshot (arXiv (Wang et al., 2019b))		63.0	80.1	84.0	86.1
Transduct.	MAML (ICML'17 (Finn et al., 2017))	RN-18	47.6 (↓ 3.8)	64.5 (↓ 5.0)	66.2 (↓ 5.7)	67.2 (↓ 3.6)
	Versa (ICLR'19 (Gordon et al., 2019))		47.8 (↓ 2.2)	61.9 (↓ 3.7)	65.6 (↓ 3.6)	67.3 (↓ 4.0)
	Entropy-min (ICLR'20 (Dhillon et al., 2020))		58.5 (↓ 5.1)	74.8 (↓ 7.3)	77.2 (↓ 8.0)	79.3 (↓ 7.9)
	LR+ICI (CVPR'2020 (Wang et al., 2020b))		58.7 (↓ 8.1)	73.5 (↓ 5.7)	78.4 (↓ 2.7)	82.1 (↓ 1.7)
	PT-MAP (arXiv (Hu et al., 2021))		60.1 (↓ 16.8)	67.1 (↓ 18.2)	68.8 (↓ 18.0)	70.4 (↓ 17.4)
	Laplacian-Shot (ICML'20 (Ziko et al., 2020))		65.4 (↓ 4.7)	81.6 (↓ 0.5)	84.1 (↓ 0.2)	86.0 (↑ 0.5)
	BD-CSPN (ECCV'20 (Liu et al., 2019b))		67.0 (↓ 2.4)	80.2(↓ 1.8)	82.9 (↓ 1.4)	84.6 (↓ 1.1)
	TIM (NeurIPS'20 (Boudiaf et al., 2020a))		67.3 (↓ 4.5)	79.8 (↓ 4.1)	82.3 (↓ 3.8)	84.2 (↓ 3.7)
	α-TIM (ours)		**67.4**	**82.5**	**85.9**	**87.9**
Induct.	Baseline (ICLR'19 (Chen et al., 2019))	WRN	62.2	81.9	85.5	87.9
	Baseline++ (ICLR'19 (Chen et al., 2019))		64.5	82.1	85.7	87.9
	Simpleshot (arXiv (Wang et al., 2019b))		66.2	82.4	85.6	87.4
Transduct.	Entropy-min (ICLR'20 (Dhillon et al., 2020))	WRN	60.4 (↓ 5.7)	76.2 (↓ 8.0)	–	–
	PT-MAP (arXiv (Hu et al., 2021))		60.6 (↓ 18.3)	66.8 (↓ 19.8)	68.5 (↓ 19.3)	69.9 (↓ 19.0)
	SIB (ICLR'20 (Hu et al., 2020))		64.7 (↓ 5.3)	72.5 (↓ 6.7)	73.6 (↓ 8.4)	74.2 (↓ 8.7)
	Laplacian-Shot (ICML'20 (Ziko et al., 2020))		68.1 (↓ 4.8)	83.2 (↓ 0.6)	85.9 (↑ 0.4)	87.2 (↑ 0.6)
	TIM (NeurIPS'20 (Boudiaf et al., 2020a))		69.8 (↓ 4.8)	81.6 (↓ 4.3)	84.2 (↓ 3.9)	85.9 (↓ 3.7)
	BD-CSPN (ECCV'20 (Liu et al., 2019b))		**70.4** (↓ 2.1)	82.3(↓ 1.4)	84.5 (↓ 1.4)	85.7 (↓ 1.1)
	α-TIM (ours)		69.8	**84.8**	**87.9**	**89.7**
			***tiered*-ImageNet**			
			1-shot	**5-shot**	**10-shot**	**20-shot**
Induct.	Baseline (ICLR'19 (Chen et al., 2019))	RN-18	63.5	83.8	87.3	89.0
	Baseline++ (ICLR'19 (Chen et al., 2019))		68.0	84.2	87.4	89.2
	Simpleshot (arXiv (Wang et al., 2019b))		69.6	84.7	87.5	89.1
Transduct.	Entropy-min (ICLR'20 (Dhillon et al., 2020))	RN-18	61.2 (↓ 5.8)	75.5 (↓ 7.6)	78.0 (↓ 7.9)	79.8 (↓ 7.9)
	PT-MAP (arXiv (Hu et al., 2021))		64.1 (↓ 18.8)	70.0 (↓ 18.8)	71.9 (↓ 17.8)	73.4 (↓ 17.1)
	LaplacianShot (ICML'20 (Ziko et al., 2020))		72.3 (↓ 4.8)	85.7 (↓ 0.5)	87.9 (↓ 0.1)	89.0 (↑ 0.3)
	BD-CSPN (ECCV'20 (Liu et al., 2019b))		74.1 (↓ 2.2)	84.8 (↓ 1.4)	86.7 (↓ 1.1)	87.9 (↓ 0.8)
	TIM (NeurIPS'20 (Boudiaf et al., 2020a))		74.1 (↓ 4.5)	84.1 (↓ 3.6)	86.0 (↓ 3.3)	87.4 (↓ 3.1)
	LR+ICI (CVPR'20 (Wang et al., 2020b))		**74.6** (↓ 6.2)	85.1 (↓ 2.8)	88.0 (↓ 2.1)	90.2 (↓ 1.2)
	α-TIM (ours)		74.4	**86.6**	**89.3**	**90.9**
Induct.	Baseline (ICLR'19 (Chen et al., 2019))	WRN	64.6	84.9	88.2	89.9
	Baseline++ (ICLR'19 (Chen et al., 2019))		68.7	85.4	88.4	90.1
	Simpleshot (arXiv (Wang et al., 2019b))		70.7	85.9	88.7	90.1
Transduct.	Entropy-min (ICLR'20 (Dhillon et al., 2020))	WRN	62.9 (↓ 6.0)	77.3 (↓ 7.5)	–	–
	PT-MAP (arXiv (Hu et al., 2021))		65.1 (↓ 19.5)	71.0 (↓ 19.0)	72.5 (↓ 18.3)	74.0 (↓ 17.7)
	LaplacianShot (ICML'20 (Ziko et al., 2020))		73.5 (↓ 5.3)	86.8 (↓ 0.5)	88.6 (↓ 0.4)	89.6 (↓ 0.2)
	BD-CSPN (ECCV'20 (Liu et al., 2019b))		75.4 (↓ 2.3)	85.9 (↓ 1.5)	87.8 (↓ 1.0)	89.1 (↓ 0.6)
	TIM (NeurIPS'20 (Boudiaf et al., 2020a))		75.8 (↓ 4.5)	85.4 (↓ 3.5)	87.3 (↓ 3.2)	88.7 (↓ 2.9)
	α-TIM (ours)		**76.0**	**87.8**	**90.4**	**91.9**

Table 3.2 Comparisons of state-of-the-art methods in our realistic setting on CUB. Query sets are sampled following a Dirichlet distribution with $\boldsymbol{a} = 2 \cdot \mathbb{1}_K$. Accuracy is averaged over 10,000 tasks. A red arrow (↓) indicates a performance drop between the artificially-balanced setting and our testing procedure, and a blue arrow (↑) an improvement. Arrows are not displayed for the inductive methods as, for these, there is no significant change in performance between both settings (expected). ‘–’ signifies the result was computationally intractable to obtain

			CUB			
			1-shot	**5-shot**	**10-shot**	**20-shot**
Induct.	Baseline (ICLR’19 (Chen *et al.*, 2019))	RN-18	64.6	86.9	90.6	92.7
	Baseline++ (ICLR’19 (Chen *et al.*, 2019))		69.4	87.5	91.0	93.2
	Simpleshot (arXiv (Wang *et al.*, 2019b))		70.6	87.5	90.6	92.2
Transduct.	PT-MAP (arXiv (Hu *et al.*, 2021))	RN-18	65.1 (↓ 20.4)	71.3 (↓ 20.0)	73.0 (↓ 19.2)	72.2 (↓ 18.9)
	Entropy-min (ICLR’20 (Dhillon *et al.*, 2020))		67.5 (↓ 5.3)	82.9 (↓ 6.0)	85.5 (↓ 5.6)	86.8 (↓ 5.7)
	Laplacian-Shot (ICML’20 (Ziko *et al.*, 2020))		73.7 (↓ 5.2)	87.7 (↓ 1.1)	89.8 (↓ 0.7)	90.6 (↓ 0.5)
	BD-CSPN (ECCV’20 (Liu *et al.*, 2019b))		74.5 (↓ 3.4)	87.1 (↓ 1.8)	89.3 (↓ 1.3)	90.3 (↓ 1.1)
	TIM (NeurIPS’20 (Boudiaf *et al.*, 2020a))		74.8 (↓ 5.5)	86.9 (↓ 3.6)	89.5 (↓ 2.9)	91.7 (↓ 2.8)
	α-TIM (ours)		**75.7**	**89.8**	**92.3**	**94.6**

supplementary material that α-TIM keeps a consistent relative improvement when evaluated under the same setting.

3.6.3 Ablation studies

In-depth comparison of TIM and α-TIM. While not included in the main Table 3.1, keeping the same hyper-parameters for TIM as prescribed in the original paper (Boudiaf *et al.*, 2020a) would result in a drastic drop of about 18% across the benchmarks. As briefly mentioned in section 3.4 and implemented for tuning (Boudiaf *et al.*, 2020a) in Table 3.1, adjusting marginal-entropy weight λ in Eq. (3.1) strongly helps in imbalanced scenarios, reducing the drop from 18% to only 4%. However, we argue that such a strategy is sub-optimal in comparison to using α-divergences, where the only hyper-parameter controlling the flexibility of the marginal-distribution term becomes α. First, as seen from Table 3.1, our α-TIM achieves consistently better performances with the same budget of hyper-parameter optimization as the standard TIM. In fact, in higher-shots scenarios (5 or higher), the performances of α-TIM are substantially better that the standard mutual information (*i.e.* TIM). Even more crucially, we show in Figure 3.3 that α-TIM does not fail drastically when α is chosen sub-optimally, as opposed to the case of weighting parameter λ for the TIM formulation. We argue that such a robustness makes of α-divergences a particularly interesting choice for more practical applications, where such a tuning might be intractable. Our results points to the high potential of α-divergences as loss functions for leveraging unlabelled data, beyond the few-shot scenario, e.g., in semi-supervised or unsupervised domain adaptation problems.

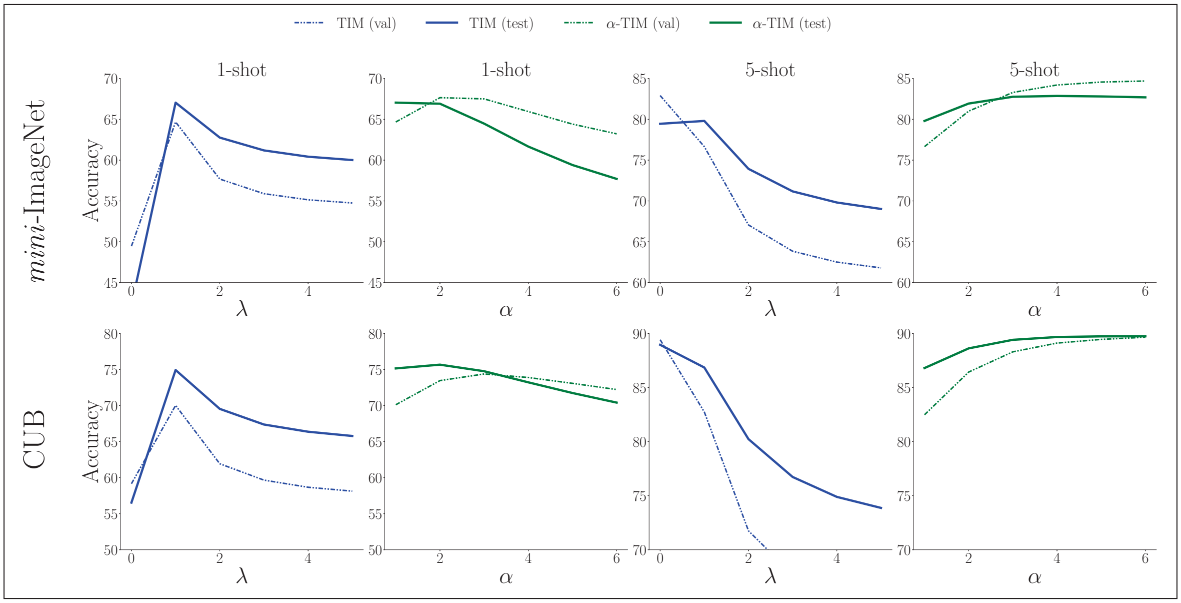

Figure 3.3 Validation and Test accuracy versus λ for TIM (Boudiaf *et al.*, 2020a) and α for our proposed α-TIM, using our protocol. Results are obtained with a RN-18. Best viewed in color

Varying imbalance severity. While our main experiments used a fixed value $\boldsymbol{a} = 2 \cdot \mathbb{1}_K$, we wonder whether our conclusions generalize to different levels of imbalance. Controlling for Dirichlet's parameter $\boldsymbol{a}$ naturally allows us to vary the imbalance severity. In Figure 3.4, we display the results obtained by varying $\boldsymbol{a}$, while keeping the same hyper-parameters obtained through our validation protocol. Generally, most methods follow the expected trend: as $\boldsymbol{a}$ decreases and tasks become more severely imbalanced, performances decrease, with sharpest losses for TIM (Boudiaf *et al.*, 2020a) and PT-MAP (Hu *et al.*, 2021). In fact, past a certain imbalance severity, the inductive baseline in (Wang *et al.*, 2019b) becomes more competitive than most transductive methods. Interestingly, both ziko2020laplacianShot (Ziko *et al.*, 2020) and our proposed α-TIM are able to cope with extreme imbalance, while still conserving good performances on balanced tasks.

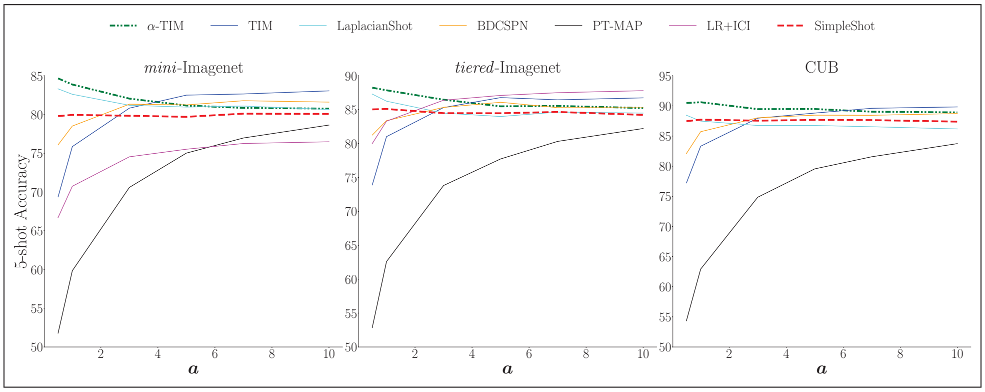

Figure 3.4 5-shot test accuracy of transductive methods versus imbalance level (lower $\boldsymbol{a}$ corresponds to more severe imbalance, as depicted in Figure 3.1)

On the use of transductive Batch-Norm. In the case of imbalanced query sets, we note that transductive batch normalization (e.g as done in the popular MAML (Finn, Xu & Levine, 2018)) hurts the performances. This aligns with recent observations from the concurrent work in (Burns & Steinhardt, 2021), where a shift in the marginal label distribution is clearly identified as a failure case of statistic alignment via batch normalization.

Conclusion

We make the unfortunate observation that recent transductive few-shot methods claiming large gains over inductive ones may perform worse when evaluated with our realistic setting. The artificial balance of the query sets in the vanilla setting makes it easy for transductive methods to implicitly encode this strong prior at meta-training stage, or even explicitly at inference. When rendering such a property obsolete at test-time, the current top-performing method becomes noncompetitive, and all the transductive methods undergo performance drops. Future works could study the mixed effect of imbalance on both support and query sets. We hope that our observations encourage the community to rethink the current transductive literature, and build upon our work to provide fairer grounds of comparison between inductive and transductive methods.

Acknowledgments

This project was supported by the Natural Sciences and Engineering Research Council of Canada (Discovery Grant RGPIN 2019-05954). This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement №792464.

CHAPTER 4

OPEN-SET LIKELIHOOD MAXIMIZATION FOR FEW-SHOT LEARNING

Malik Boudiaf[1*] , Etienne Bennequin[2*] , Myriam Tami[3] , Antoine Toubhans[2] ,
Pablo Piantanida[3] , Celine Hudelot[3] , Ismail Ben Ayed[1]

[1] ÉTS Montréal, QC, Canada,
[2] Sicara, France,
[3] CentraleSupelec-Université Paris-Saclay, France

[*] Equal contributions

Poster at the Conference on Computer Vision and Pattern Recognition (CVPR) 2023,
Vancouver.

Abstract

We tackle the Few-Shot Open-Set Recognition (FSOSR) problem, i.e. classifying instances among a set of classes for which we only have a few labeled samples, while simultaneously detecting instances that do not belong to any known class. We explore the popular transductive setting, which leverages the unlabelled query instances at inference. Motivated by the observation that existing transductive methods perform poorly in open-set scenarios, we propose a generalization of the maximum likelihood principle, in which latent scores down-weighing the influence of potential outliers are introduced alongside the usual parametric model. Our formulation embeds supervision constraints from the support set and additional penalties discouraging overconfident predictions on the query set. We proceed with a block-coordinate descent, with the latent scores and parametric model co-optimized alternately, thereby benefiting from each other. We call our resulting formulation *Open-Set Likelihood Optimization* (OSLO). OSLO is interpretable and fully modular; it can be applied on top of any pre-trained model seamlessly. Through extensive experiments, we show that our method surpasses existing inductive and transductive methods on both aspects of open-set recognition, namely inlier classification and outlier detection. Code is available as part of the supplementary material.

4.1 Introduction

Few-shot classification consists in recognizing concepts for which we have only a handful of labeled examples. These form the *support set*, which, together with a batch of unlabeled instances (the *query set*), constitute a few-shot task. Most few-shot methods classify the unlabeled query samples of a given task based on their similarity to the support instances in some feature space (Snell *et al.*, 2017). This implicitly assumes a *closed-set* setting for each task, i.e. query instances are supposed to be constrained to the set of classes explicitly defined by the support set. However, the real world is open and this closed-set assumption may not hold in practice, especially for limited support sets. Whether they are unexpected items circulating on an assembly line, a new dress not yet included in a marketplace's catalog, or a previously undiscovered species of fungi, *open-set instances* occur everywhere. When they do, a closed-set classifier will falsely label them as the closest known class.

This drove the research community toward open-set recognition i.e. recognizing instances with the awareness that they may belong to unknown classes. In the large-scale settings, the literature abounds of methods designed specifically to detect open-set instances while maintaining good accuracy on closed-set instances (Scheirer, de Rezende Rocha, Sapkota & Boult, 2012; Bendale & Boult, 2016; Zhou, Ye & Zhan, 2021). Very recently, the authors of (Liu, Kang, Li, Hua & Vasconcelos, 2020b) introduced a Few-Shot Open-Set Recognition (FSOSR) setting, in which query instances may not belong to any known class. The study in (Liu *et al.*, 2020b), together with other recent follow-up works (Jeong, Choi & Kim, 2021; Huang, Ma, Han & Chang, 2022), exposed FSOSR to be a difficult task in the inductive setting. On the other hand, owing to its staggering performance in settings where data is scarce (Vapnik, 2013), the transductive setting has become a prominent research direction for few-shot classification and the subject of a large body of recent few-shot works, e.g. (Veilleux, Boudiaf, Piantanida & Ben Ayed, 2021; Dhillon *et al.*, 2020; Liu *et al.*, 2020c; Ziko *et al.*, 2020; Boudiaf *et al.*, 2020a; Wang *et al.*, 2020b; Hu *et al.*, 2021; Boudiaf *et al.*, 2021), among many others. Indeed, transductive few-shot methods make joint predictions for the query samples of each given few-shot task, rather than one sample at a time as in the inductive setting (Snell *et al.*, 2017). By leveraging the statistics of the query set, transductive methods yield performances that are substantially better than their inductive counterparts (Boudiaf *et al.*, 2020a; Veilleux *et al.*, 2021). Yet we observe that current transductive few-shot methods are significantly worse than simple inductive baselines at detecting outliers. This might explain why transduction, despite its popularity in few-shot learning, had not yet been explored in the FSOSR context. Indeed, the presence of outliers in the unlabelled data tends to conflict with existing transductive principles. Specifically, we show in our experiments that transductive methods tend to match the classification confidence for open-set instances with that of closed-set instances, making prediction-based detection more difficult.

Contributions. In this work, we aim at designing a principled framework that reconciles transduction with the open nature of the FSOSR problem. Our idea is simple but powerful: instead of finding heuristics to assess the *outlierness* of each unlabelled query sample, we treat this score as a latent variable of the problem. Based on this idea, we propose a generalization of the maximum likelihood principle, in which the introduced latent scores weigh potential outliers down, thereby preventing the parametric model from fitting those samples. Our generalization embeds additional supervision constraints from the support set and penalties discouraging overconfident predictions. We proceed with a block-coordinate descent optimization of our objective, with the closed-set soft assignments, *outlierness* scores, and parametric models co-optimized alternately, thereby benefiting from each other. We call our resulting formulation *Open-Set Likelihood Optimization* (OSLO). OSLO provides highly interpretable and closed-form solutions within each iteration for both the soft assignments, *outlierness* variables, and the parametric model. Additionally, OSLO is fully modular; it can be applied on top of any pre-trained model seamlessly.

Empirically, we show that OSLO significantly surpasses its inductive and transductive competitors alike for both outlier detection and closed-set prediction. Applied on a wide variety of architectures and training strategies and without any re-optimization of its parameters, OSLO's improvement over a strong baseline remains large and consistent. This modularity allows our method to fully benefit from the latest advances in standard image recognition. Before diving into the core content, let us summarize our contributions:

1. To the best of our knowledge, we realize the first study and benchmarking of transductive methods for the Few-Shot Open-Set Recognition setting. We reproduce and benchmark five state-of-the-art transductive methods.
2. We introduce Open-Set Likelihood Optimization (OSLO), a principled extension of the Maximum Likelihood framework that explicitly models and handles the presence of outliers. OSLO is interpretable and modular i.e. can be applied on top of any pre-trained model seamlessly.
3. Through extensive experiments spanning five datasets and a dozen of pre-trained models, we show that OSLO consistently surpasses both inductive and existing transductive methods in detecting open-set instances while competing with the strongest transductive methods in classifying closed-set instances.

4.2 Related Works

Few-shot classification (FSC) methods. Many FSC works involve episodic training (Vinyals, Blundell, Lillicrap, Kavukcuoglu & Wierstra, 2016a), in which a neural network acting as a feature extractor is trained on artificial tasks sampled from the training set. This replication of the inference scenario during training is intended to make the learned representation more robust to new classes. However, several recent works have shown that simple fine-tuning baselines are competitive in comparison to sophisticated episodic methods, e.g. (Chen *et al.*, 2019; Goldblum *et al.*, 2020), motivating a new direction of few-shot learning research towards the development of model-agnostic methods that do not involve any specific training strategy (Dhillon *et al.*, 2020).

Transductive FSC. Transductive FSC methods leverage statistics of the query set as unlabeled data to improve performance, through model fine-tuning (Dhillon *et al.*, 2020), Laplacian regularization (Ziko *et al.*, 2020), clustering (Lichtenstein *et al.*, 2020), mutual information maximization (Boudiaf *et al.*, 2020a; Veilleux *et al.*, 2021), prototype rectification (Liu *et al.*, 2020c), or optimal transport (Bennequin, Bouvier, Tami, Toubhans & Hudelot, 2021; Hu *et al.*, 2021; Lazarou, Stathaki & Avrithis, 2021), among other transduction strategies. The idea of maximizing the likelihood of both support and query samples under a model parameterized by class prototypes is proposed by (Yang, Liu, Li, Jiao & Ye, 2020b) for few-shot segmentation. However, their method relies on the closed-set assumption. Differing from previous works, our framework leverages an additional latent variable, the *inlierness* score.

Open-set recognition (OSR). OSR aims to enable classifiers to detect instances from unknown classes (Scheirer *et al.*, 2012). Prior works address this problem in the large-scale setting by augmenting the SoftMax activation to account for the possibility of unseen classes (Bendale & Boult, 2016), generating artificial outliers (Ge, Demyanov, Chen & Garnavi, 2017; Neal, Olson, Fern, Wong & Li, 2018), improving closed-set accuracy (Vaze, Han, Vedaldi & Zisserman, 2022), or using placeholders to anticipate novel classes' distributions with adaptive decision boundaries (Zhou *et al.*, 2021). All these methods involve the training of deep neural networks on a specific class set. Therefore, they are not fully fit for the few-shot setting. In this work, we use simple yet effective adaptations of OpenMax (Bendale & Boult, 2016) and PROSER (Zhou *et al.*, 2021) as strong baselines for FSOSR.

Few-shot open-set recognition. In the few-shot setting, methods must detect open-set instances while only a few closed-set instances are available. (Liu *et al.*, 2020b) use meta-learning on pseudo-open-set tasks to train a model to maximize the classification entropy of open-set instances. (Jeong *et al.*, 2021) use transformation consistency to measure the divergence between a query image and the set of class prototypes. (Huang *et al.*, 2022) use an attention mechanism to generate a negative prototype for outliers. These methods require the optimization of a separate model with a specific episodic training strategy. Nonetheless, as we show in section 4.5, they bring marginal improvement over simple adaptations of standard OSR methods to the few-shot setting. In comparison, our method doesn't require any specific training and can be plugged into any feature extractor without further optimization.

4.3 Few-Shot Open-Set Recognition

Model training. Let us denote the raw image space $\mathcal{X}$. As per the standard Few-Shot setting, we assume access to a *base* dataset $\mathcal{D}_{\text{base}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1\dots|\mathcal{D}_{\text{base}}|}$ with base classes $\mathcal{Y}_{\text{base}}$, such that $\boldsymbol{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}_{\text{base}}$. We use $\mathcal{D}_{\text{base}}$ to train a feature extractor $f_{\boldsymbol{\phi}}$. Our method developed later in section 4.4, freezes $f_{\boldsymbol{\phi}}$ and performs inference directly on top of the extracted features for each task.

Testing. Given a set of *novel* classes $\mathcal{Y}_{\text{novel}}$ disjoint from base classes i.e. $\mathcal{Y}_{\text{novel}} \cap \mathcal{Y}_{\text{base}} = \emptyset$, a K-way FSOSR task is formed by sampling a set of K *closed-set* classes $\mathcal{Y}_{\text{closed-set}} \subset \mathcal{Y}_{\text{novel}}$, a support set of labeled instances $\mathbb{S} = \{(\boldsymbol{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}_{\text{closed-set}}\}_{i=1}^{|\mathbb{S}|}$ and a query set $\mathbb{Q} = \{\boldsymbol{x}_i \in \mathcal{X}\}_{i=|\mathbb{S}|+1}^{|\mathbb{S}|+|\mathbb{Q}|}$. In the standard few-shot setting, the unknown ground-truth query labels $\{y_i\}_{i=|\mathbb{S}|+1}^{|\mathbb{S}|+|\mathbb{Q}|}$ are assumed to be restricted to closed-set classes i.e. $\forall i,\ y_i \in \mathcal{Y}_{\text{closed-set}}$. In FSOSR, however, query labels may also belong to an additional set $\mathcal{Y}_{\text{OS}} \subset \mathcal{Y}_{\text{novel}}$ of *open-set* classes i.e. $\forall\ i > |\mathbb{S}|,\ y_i \in \mathcal{Y}_{\text{closed-set}} \cup \mathcal{Y}_{\text{OS}}$ with $\mathcal{Y}_{\text{closed-set}} \cap \mathcal{Y}_{\text{OS}} = \emptyset$. For easy referencing, we refer to query samples from the closed-set classes $\mathcal{Y}_{\text{closed-set}}$ as *inliers* and to query samples from open-set classes $\mathcal{Y}_{\text{OS}}$ as *outliers*. For each query image $\boldsymbol{x}_i$, the goal of FSOSR is to simultaneously assign a closed-set prediction and an *outlierness* (or equivalently *inlierness*) score.

Transductive FSOSR. As a growing part of the Few-Shot literature, Transductive Few-Shot Learning assumes that unlabelled samples from the query set are observed at once, such that the structure of unlabelled data can be leveraged to help constrain ambiguous few-shot tasks. Transductive methods have achieved impressive improvements over inductive methods in standard closed-set FSC (Dhillon *et al.*, 2020; Boudiaf *et al.*, 2020a; Ziko *et al.*, 2020; Hu *et al.*, 2021). We expect that transductive methods can help us improve overall open-set performance. While we find this to generally hold for closed-set predictive performance, we empirically show in section 4.5 that accuracy gains systematically come along significant outlier detection degradation, indicating that transductive methods are not equipped to handle open-set recognition. In the following, we take up the challenge of designing a transductive optimization framework that leverages the presence of outliers to improve its performance.

4.4 Open-Set Likelihood

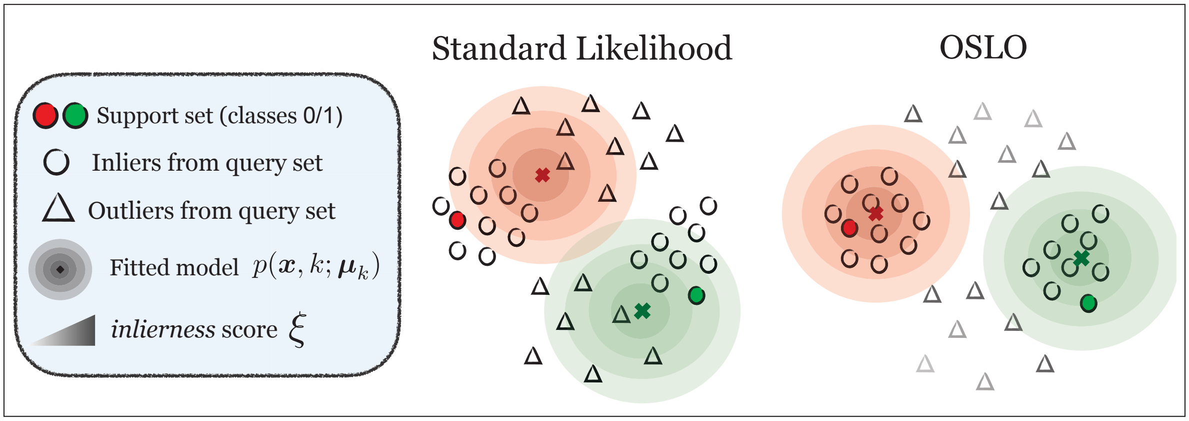

Figure 4.1 **Intuition behind OSLO.** Standard transductive likelihood (left) tries to enforce high likelihood for all points, including outliers. OSLO (right) instead treats the *outlierness* of each sample as a latent variable to be solved alongside the parametric model. Besides yielding a principled *outlierness* score for open-set detection, it also allows the fitted parametric model to effectively disregard samples deemed outliers, and therefore provide a better approximation of underlying class-conditional distributions

In this section, we introduce OSLO, a novel extension of the standard likelihood designed for transductive FSOSR. Unlike existing transductive methods, OLSO explicitly models and handles the potential presence of outliers, which allows it to outperform inductive baselines on both aspects of the open-set scenario.

Observed variables. We start by establishing the observed variables of the problem. As per the traditional setting, we observe images from the support set $\{\boldsymbol{x}_i\}_{i=1}^{|\mathbb{S}|}$ and their associated labels $\{y_i\}_{i=1}^{|\mathbb{S}|}$. The transductive setting also allows us to observe images from the query set. For notation convenience, we concatenate all images in $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|}$.

Latent variables. Our goal is to predict the class of each sample in the query set $\mathbb{Q}$, as well as their *inlierness*, i.e. the model's belief in a sample being an inlier or not. This naturally leads us to consider latent class assignments $z_i \in \Delta^K$ describing the membership of sample i to each closed-set class, with $\Delta^K = \{z \in [0,1]^K : z^T\mathbf{1} = 1\}$ the K-dimensional simplex. Additionally, we consider latent *inlierness* scores $\xi_i \in [0,1]$ close to 1 if the model considers sample i as an inlier. For notation convenience, we consider latent assignments and *inlierness* scores for all samples, including those from the support, and group everything in $\boldsymbol{Z} = \{\boldsymbol{z}_i\}_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|}$ and $\boldsymbol{\xi} = \{\xi_i\}_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|}$. Note that support samples are inliers, and we know their class. Therefore $\forall i \leq |\mathbb{S}|$, the constraints $\boldsymbol{z}_i = \boldsymbol{y}_i$ and $\xi_i = 1$ will be imposed, where $\boldsymbol{y}_i$ is the one-hot encoded version of y_i

Parametric model. The final ingredient we need to formulate is a parametric joint model over observed features and assignments. Following standard practice, we model the joint distribution as a balanced mixture of standard Gaussian distributions, parameterized by the centroids $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$:

$$p(\boldsymbol{x}, k; \boldsymbol{\mu}) = p(k)p(\boldsymbol{x}|k) \propto \exp(-\frac{1}{2}\left\|f_{\boldsymbol{\phi}}(\boldsymbol{x}) - \boldsymbol{\mu}_k\right\|^2) \tag{4.1}$$

Note that we're using a uniform assumption, which could hurt performance in cases where the task of interest exhibits severe class imbalance. Although the goal of this Chapter is to evaluate the influence of outliers, techniques developed in Chapters 3 and 5, namely switching to α-divergence or treating $\{p(k)\}_{k=1}^K$ as latent variables, could also be applied to address the problem of class imbalance.

Objective. Using the i.i.d. assumption, we start by writing the standard likelihood objective:

$$p(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\mu}) = \prod_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|} \prod_{k=1}^{K} p(\boldsymbol{x}_i, k; \boldsymbol{\mu})^{z_{ik}} \tag{4.2}$$

Without loss of generality, we consider the log-likelihood:

$$\log(p(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\mu})) = \sum_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|} \sum_{k=1}^{K} z_{ik} \log(p(\boldsymbol{x}_i, k; \boldsymbol{\mu})) \tag{4.3}$$

Eq. (4.2) tries to enforce a high likelihood of all samples under our parametric model p. This becomes sub-optimal in the presence of outliers, which should ideally be disregarded. Figure 4.1 illustrates this phenomenon on a toy 2D drawing. To downplay this issue, we introduce *Open-Set Likelihood Optimization* (OSLO), a generalization of the standard likelihood framework, which leverages latent *inlierness* scores to weigh samples:

$$\mathcal{L}_{\text{OSLO}}(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\xi}; \boldsymbol{\mu}) = \sum_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|} \xi_i \sum_{k=1}^{K} z_{ik} \log\left(p(\boldsymbol{x}_i, k; \boldsymbol{\mu})\right) \tag{4.4}$$

Eq. (4.4) can be interpreted as follows: samples believed to be inliers i.e. $\xi_i \approx 1$ will be required to have a high likelihood under our model p, whereas outliers won't. Note that $\boldsymbol{\xi}$ is treated as a variable of optimization, and is co-optimized alongside $\boldsymbol{\mu}$ and $\boldsymbol{Z}$. Finally, to prevent overconfident latent scores, we consider a *penalty* term on both $\boldsymbol{Z}$ and $\boldsymbol{\xi}$:

$$\mathcal{L}_{\text{soft}} = \sum_{i=|\mathbb{S}|+1}^{|\mathbb{S}|+|\mathbb{Q}|} \lambda_z \mathcal{H}(\boldsymbol{z}_i) + \lambda_\xi \mathcal{H}(\boldsymbol{\xi}_i) \tag{4.5}$$

where $\boldsymbol{\xi}_i = [1 - \xi_i, \xi_i]$, and $\mathcal{H}(\boldsymbol{p}) = -\boldsymbol{p}^\top \log(\boldsymbol{p})$ denotes the entropy, which encourages smoother assignments.

Link to entropy minimization. The standard likelihood objective Eq. 4.2, complemented with the regularization (negative entropy) term in Eq. 4.5 is highly related to the principle of entropy minimization used throughout this thesis, and specifically in Chapter 2. Indeed, using the chain rule $p(\boldsymbol{x}, k) = p(\boldsymbol{x})p(k|\boldsymbol{x}) \propto p(\boldsymbol{x})$ and solving for z_{ik} would yield $z_{ik} = p(k|\boldsymbol{x}_i; \boldsymbol{\mu})$. Plugging that back in 4.2 yields the conditional entropy term $\mathcal{H}(Y|X)$ used in 2.

Optimization. We are now ready to formulate OSLO's optimization problem:

$$\begin{aligned} \max_{\boldsymbol{\mu}, \boldsymbol{Z}, \boldsymbol{\xi}} \quad & \mathcal{L}_{\text{OSLO}}(\boldsymbol{Z}, \boldsymbol{\xi}, \boldsymbol{\mu}) - \mathcal{L}_{\text{soft}}(\boldsymbol{Z}, \boldsymbol{\xi}) \\ \text{s.t} \quad & \boldsymbol{z}_i \in \Delta^K, \quad \xi_i \in [0,1] \quad \forall\, i \\ & \boldsymbol{z}_i = \boldsymbol{y}_i, \quad \xi_i = 1, \quad i \leq |\mathbb{S}| \end{aligned} \tag{4.6}$$

Problem (4.6) is strictly convex with respect to each variable when the other variables are fixed. Therefore, we proceed with a block-coordinate ascent, which alternates three iterative steps, each corresponding to a closed-form solution for one of the variables.

Proposition 4.4.0.1. *OSLO's optimization problem* (4.6) *can be minimized by alternating the following updates:*

$$\xi_i^{(t+1)} = \begin{cases} 1 & \textit{if}\, i \leq |\mathbb{S}| \\ \sigma\left(\dfrac{1}{\lambda_\xi} \displaystyle\sum_{k=1}^{K} z_{ik}^{(t)} \log p(\boldsymbol{x}_i, k; \boldsymbol{\mu}^{(t)}))\right) & \textit{else} \end{cases}$$

$$\boldsymbol{z}_i^{(t+1)} \propto \begin{cases} \boldsymbol{y}_i & \textit{if}\, i \leq |\mathbb{S}| \\ \exp\left(\dfrac{\xi_i^{(t+1)}}{\lambda_z} \log p(\boldsymbol{x}_i, \cdot\, ; \boldsymbol{\mu}^{(t)})\right) & \textit{else} \end{cases}$$

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{1}{\displaystyle\sum_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|} \xi_i^{(t+1)} z_{ik}^{(t+1)}} \sum_{i=1}^{|\mathbb{S}|+|\mathbb{Q}|} \xi_i^{(t+1)} z_{ik}^{(t+1)} f_{\boldsymbol{\phi}}(\boldsymbol{x}_i)$$

where σ denotes the sigmoid operation.

The proof of proposition 4.4.0.1 is performed by derivation of $\mathcal{L}_{\text{OSLO}}(\boldsymbol{Z}, \boldsymbol{\xi}, \boldsymbol{\mu}) + \mathcal{L}_{\text{soft}}(\boldsymbol{Z}, \boldsymbol{\xi})$ and deferred to the supplementary material. The optimal solution for the *inlierness* score ξ_i appears very intuitive, and essentially conveys that samples with high likelihood under the current parametric model should be considered inliers. We emphasize that **beyond providing a principled *outlierness score*, as $1 - \xi_i$, the presence of ξ_i allows to refine and improve the closed-set parametric model**. In particular, ξ_i acts as a sample-wise temperature in the

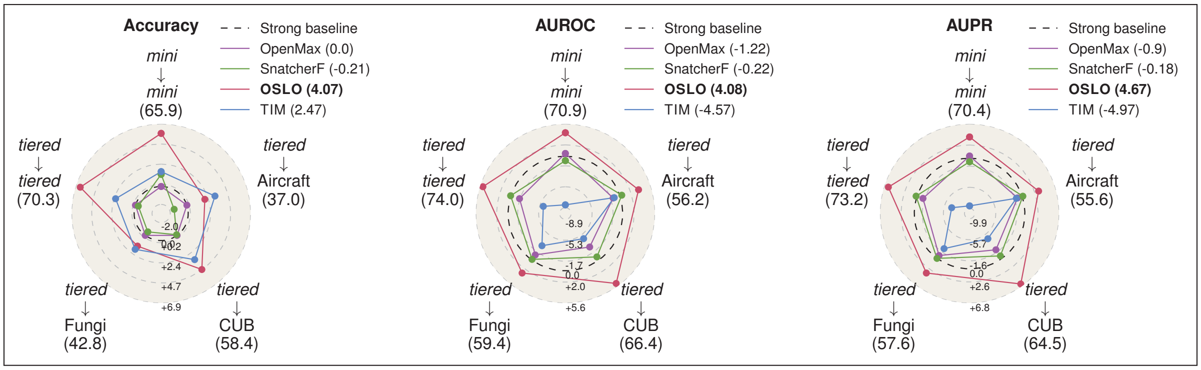

Figure 4.2 **OSLO improves open-set performances on a wide variety of tasks.** Relative 1-shot performance of the best methods of each family w.r.t the *Strong baseline* using a ResNet-12, across a set of 5 scenarios, including 3 with domain-shift. Each vertex represents one scenario, e.g. *tiered*→Fungi (x) means the feature extractor was pre-trained on *tiered*-ImageNet, test tasks are sampled from Fungi, and the *Strong Baseline* performance is x. For each method, the average relative improvement across the 5 scenarios is reported in parenthesis in the legend. The same charts are provided in the supplementary materials for the 5shot setting and using a WideResNet backbone

update of z_i, encouraging outliers ($\xi_i \approx 0$) to have a uniform distribution over closed-set classes. Additionally, those samples contribute less to the update of closed-set prototypes $\boldsymbol{\mu}$, as each sample's contribution is weighted by ξ_i.

4.5 Experiments

4.5.1 Experimental setup

Baselines. One goal of this work is to fairly evaluate different strategies to address the FSOSR problem. In particular, we benchmark 4 families of methods: (i) popular Outlier Detection methods, e.g. Nearest-Neighbor (Ramaswamy, Rastogi & Shim, 2000), (ii) Inductive Few-Shot classifiers, e.g. SimpleShot (Wang *et al.*, 2019b) (iii) Inductive Open-Set methods formed by standard methods such as OpenMax (Bendale & Boult, 2016) and Few-Shot methods such as Snatcher (Jeong *et al.*, 2021) (iv) Transductive classifiers, e.g. TIM (Boudiaf *et al.*, 2020a), that implicitly rely on the closed-set assumption, and finally (v) Transductive Open-Set introduced in this work through OSLO. Following (Jeong *et al.*, 2021), closed-set few-shot classifiers are

Table 4.1 **Standard Benchmarking**. Evaluating different families of methods on the FSOSR problem on *mini*-ImageNet and *tiered*-ImageNet using a ResNet-12. For each column, a light-gray standard deviation is indicated, corresponding to the maximum deviation observed across methods for that metric. Best methods are shown in bold. Results marked with ⋆ are reported from their original paper

Strategy	Method	***mini*-ImageNet**							
		1-shot				5-shot			
		Acc	AUROC	AUPR	Prec@0.9	Acc	AUROC	AUPR	Prec@0.9
		±0.72	±0.79	±0.69	±0.47	±0.44	±0.73	±0.61	±0.56
OOD detection	*k*-NN	-	70.86	70.43	58.23	-	76.22	76.36	61.48
	IForest	-	55.59	55.24	52.18	-	62.80	61.62	54.77
	OCVSM	-	69.67	69.71	57.35	-	68.49	65.60	59.24
	PCA	-	67.23	66.50	56.67	-	75.24	75.53	60.73
	COPOD	-	50.60	51.85	50.92	-	51.63	52.65	51.31
	HBOS	-	58.26	57.41	53.06	-	61.11	60.18	54.30
Inductive classifiers	SimpleShot (Wang *et al.*, 2019b)	65.90	64.99	63.78	55.77	81.72	70.61	70.06	57.91
	Baseline ++ (Chen *et al.*, 2019)	65.81	65.15	63.85	55.87	81.86	66.37	65.58	56.33
	FEAT (Ye *et al.*, 2020b)	67.23	52.45	54.44	50.00	82.00	53.25	56.48	50.00
Inductive Open-Set	PEELER⋆ (Liu *et al.*, 2020b)	65.86	60.57	-	-	80.61	67.35	-	-
	TANE-G⋆ (Huang *et al.*, 2022)	68.11	72.41	-	-	83.12	79.85	-	-
	SnatcherF (Jeong *et al.*, 2021)	67.23	70.10	69.74	58.02	82.00	76.57	76.97	61.64
	OpenMax (Bendale & Boult, 2016)	65.90	71.34	70.86	58.67	82.23	77.42	77.63	62.35
	PROSER (Zhou *et al.*, 2021)	65.00	68.93	68.84	57.03	80.08	74.98	75.58	60.11
Transductive classifiers	LaplacianShot (Ziko *et al.*, 2020)	70.59	53.13	54.59	52.06	82.94	57.17	57.90	52.56
	BDCSPN (Liu *et al.*, 2020c)	69.35	57.95	58.58	52.71	82.66	61.27	62.17	53.26
	TIM-GD (Boudiaf *et al.*, 2020a)	67.53	62.46	61.05	54.83	82.49	67.19	66.15	56.70
	PT-MAP (Hu *et al.*, 2021)	66.32	59.05	58.67	53.74	78.12	62.78	62.48	54.67
	LR-ICI (Wang *et al.*, 2020b)	68.24	49.96	51.61	50.45	81.77	51.82	53.49	50.80
Transductive Open-Set	OSLO (ours)	**71.73**	**74.92**	**74.61**	**60.95**	**83.40**	**82.59**	**82.34**	**66.98**
		***tiered*-ImageNet**							
		1-shot				5-shot			
		Acc	AUROC	AUPR	Prec@0.9	Acc	AUROC	AUPR	Prec@0.9
		±0.74	±0.76	±0.71	±0.52	±0.52	±0.68	±0.75	±0.57
OOD detection	*k*-NN	-	73.97	73.15	60.74	-	80.22	80.06	65.47
	IForest	-	54.57	54.24	51.85	-	62.31	60.82	54.72
	OCVSM	-	71.22	71.17	58.81	-	71.20	68.23	61.09
	PCA	-	68.30	67.02	57.66	-	76.26	76.41	61.81
	COPOD	-	50.87	51.95	51.07	-	52.62	53.48	51.44
	HBOS	-	57.54	56.67	52.98	-	60.91	59.95	54.15
Inductive classifiers	SimpleShot (Wang *et al.*, 2019b)	70.27	69.78	67.89	58.54	84.94	77.38	76.28	63.21
	Baseline ++ (Chen *et al.*, 2019)	70.21	69.73	67.80	58.50	85.10	73.77	72.39	61.05
	FEAT (Ye *et al.*, 2020b)	69.94	52.49	56.74	50.00	83.96	53.30	59.81	50.00
Inductive Open-Set	PEELER⋆ (Liu *et al.*, 2020b)	69.51	65.20	-	-	84.10	73.27	-	-
	TANE-G⋆ (Huang *et al.*, 2022)	70.58	73.53	-	-	85.38	81.54	-	-
	SnatcherF (Jeong *et al.*, 2021)	69.94	74.02	73.33	60.79	83.96	81.90	81.67	66.89
	OpenMax (Bendale & Boult, 2016)	70.27	72.40	71.91	59.91	85.79	77.91	78.42	63.07
	PROSER (Zhou *et al.*, 2021)	68.48	70.07	69.87	57.99	83.34	75.84	76.56	61.12
Transductive classifiers	LaplacianShot (Ziko *et al.*, 2020)	75.66	57.82	58.41	53.67	86.23	63.75	63.65	55.36
	BDCSPN (Liu *et al.*, 2020c)	74.07	62.13	61.84	54.53	85.65	67.41	67.57	56.30
	TIM-GD (Boudiaf *et al.*, 2020a)	72.56	68.08	65.97	57.84	85.70	74.67	73.06	61.59
	PT-MAP (Hu *et al.*, 2021)	71.13	64.48	62.94	56.25	82.81	71.08	69.89	59.11
	LR-ICI (Wang *et al.*, 2020b)	73.80	49.32	51.41	50.35	85.21	51.65	53.85	50.79
Transductive Open-Set	OSLO (ours)	**76.64**	**79.06**	**79.07**	**64.36**	**86.35**	**86.92**	**87.28**	**71.98**

turned into open-set classifiers by considering the negative of the maximum probability as a measure of outlierness. Furthermore, we found that applying a center-normalize transformation $\psi_{\boldsymbol{v}} : \boldsymbol{x} \mapsto (\boldsymbol{x} - \boldsymbol{v})/||\boldsymbol{x} - \boldsymbol{v}||_2$ on the features extracted by $f_{\boldsymbol{\phi}}$ benefited all methods. Therefore, we apply it to the features before applying any method, using an inductive *Base centering* (Wang *et al.*, 2019b) for inductive methods $\boldsymbol{v}_{Base} = \frac{1}{|\mathcal{D}_{base}|} \sum_{\boldsymbol{x} \in \mathcal{D}_{base}} f_{\boldsymbol{\phi}}(x)$, and a transductive *Task centering* (Hu *et al.*, 2021) $\boldsymbol{v}_{Task} = \frac{1}{|\mathbb{S} \cup \mathbb{Q}|} \sum_{\boldsymbol{x} \in \mathbb{S} \cup \mathbb{Q}} f_{\boldsymbol{\phi}}(x)$ for all transductive methods. Since features are normalized, we empirically found it beneficial to re-normalize centroids $\boldsymbol{\mu}_k \leftarrow \boldsymbol{\mu}_k / ||\boldsymbol{\mu}_k||_2$ after each update from Prop. 4.4.0.1, which we show in the Appendix remains a valid minimizer of Equation 4.6 when adding the constraint $||\boldsymbol{\mu}_k||_2 = 1$.

Hyperparameters. For all methods, we define a grid over salient hyper-parameters and tune over the validation split of *mini*-ImageNet. To avoid cumbersome per-dataset tuning, and evaluate the generalizability of methods, we then keep hyper-parameters fixed across all other experiments.

Architectures and checkpoints. To provide the fairest comparison, all non-episodic methods are tuned and tested using off-the-shelf pre-trained checkpoints. All results except Figure 4.4 are produced using the pre-trained ResNet-12 and Wide-ResNet 28-10 checkpoints provided by the authors from (Ye *et al.*, 2020b). As for episodically-finetuned models required by Snatcher (Jeong *et al.*, 2021) and FEAT (Ye *et al.*, 2020b), checkpoints are obtained from the authors' respective repositories. Finally, to challenge the model-agnosticity of our method, we resort to an additional set of 10 ImageNet pre-trained models covering three distinct architectures: ResNet-50 (He *et al.*, 2016) for CNNs, ViT-B/16 (Dosovitskiy *et al.*, 2021) for vision transformers, and Mixer-B/16 (Tolstikhin *et al.*, 2021) for MLP-Mixer. Most models used are taken from the excellent TIMM library (Wightman, 2019).

Datasets and tasks. We experiment with a total of 5 vision datasets. As standard FSC benchmarks, we use the *mini*-ImageNet (Vinyals *et al.*, 2016a) dataset with 100 classes and the larger *tiered*-ImageNet (Ren *et al.*, 2018) dataset with 608 classes. We also experiment on more challenging cross-domain tasks formed by using 3 finer-grained datasets: the Caltech-UCSD Birds 200 (Welinder *et al.*, 2010) (CUB) dataset, with 200 classes, the FGVC-Aircraft dataset (Maji, Rahtu, Kannala, Blaschko & Vedaldi, 2013) with 100 classes, and the Fungi classification challenge (Schroeder & Cui, 2018) with 1394 classes. Following standard FSOSR protocol, support sets contain $|\mathcal{Y}_{\text{closed-set}}| = 5$ closed-set classes with 1 or 5 instances, or *shots*, per class, and query sets are formed by sampling 15 instances per class, from a total of ten classes: the five closed-set classes and an additional set of $|\mathcal{Y}_{\text{OS}}| = 5$ open-set classes. We follow this setting for a fair comparison with previous works (Jeong *et al.*, 2021; Liu *et al.*, 2020b) which sample open-set query instances from only 5 classes. We also report results in supplementary materials for a more general setting in which open-set query instances are sampled indifferently from all remaining classes in the test set.

4.5.2 Results

Simplest inductive methods are competitive. The first surprising result comes from analyzing the performances of standard OOD detectors on the FSOSR problem. Table 4.1 shows that k-NN and PCA outperform, by far, arguably more advanced methods that are OCVSM and Isolation Forest. This result contrasts with standard high-dimensional benchmarks (Zhao, Nasrullah & Li, 2019) where k-NN falls typically short of the latter, indicating that the very difficult challenge posed by FSOSR may lead advanced methods to overfit. In fact, Figure 4.2 shows that across 5 scenarios, the combination SimpleShot (Wang *et al.*, 2019b)+ k-NN (Ramaswamy *et al.*, 2000) formed by the simplest FS-inductive classifier and the simplest inductive OOD detector is a strong baseline that outperforms all specialized open-set methods. We refer to this combination as *Strong baseline* in Figures 4.2 and 4.4. Additional results for the Wide-ResNet architecture are provided in the supplementary material.

Transductive methods still improve accuracy but degrade outlier detection. As shown in Table 4.1, most transductive classifiers still offer a significant boost in closed-set accuracy, even in the presence of outliers in the query set. Note that this contrasts with findings from the semi-supervised literature, where standard methods drop below the baseline in the presence of even a small fraction of outliers (Yu, Ikami, Irie & Aizawa, 2020; Chen, Zhu, Li & Gong, 2020b; Saito, Kim & Saenko, 2021; Killamsetty, Zhao, Chen & Iyer, 2021). We hypothesize that the deliberate under-parametrization of few-shot methods –typically only training a linear classifier–, required to avoid overfitting the support set, partly explains such robustness. However,

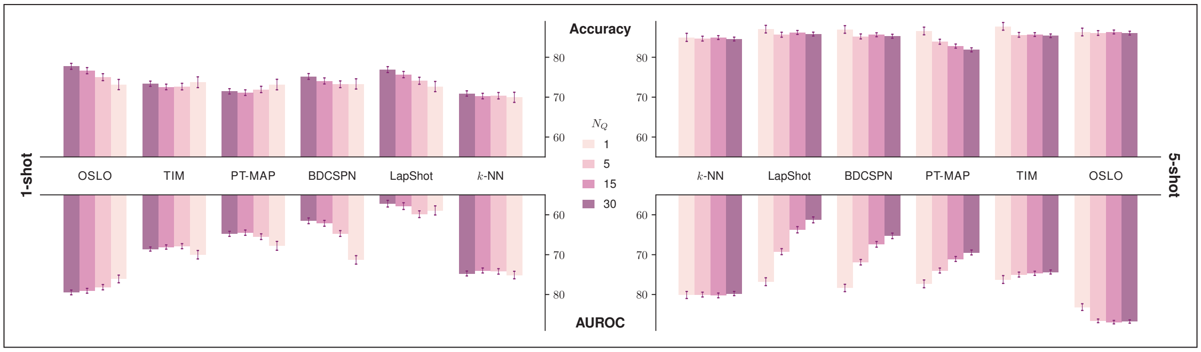

Figure 4.3 **OSLO improves performance even with few queries.** We study the closed-set (accuracy) and open-set (AUROC) performance of transductive methods depending on the size of the query set on *tiered*-ImageNet in the 1-shot and 5-shot settings. The total size $|Q|$ of the query set is obtained by multiplying the number of queries per class N_Q by the number of classes in the task (i.e. 5) and adding as many outlier queries e.g. $N_Q = 1$ corresponds to 1 query per class and 5 open-set queries i.e. $|Q| = 10$. We add the inductive method k-NN + SimpleShot to compare with a method that is by nature independent of the number of queries. The results for *mini*-ImageNet are provided in the supplementary materials

transductive methods still largely underperform in outlier detection, with AUROCs as low as 52 % (50% being a random detector) for LaplacianShot. Note that the *outlierness* score for these methods is based on the negative of the maximum probability, therefore this result can be interpreted as transductive methods having artificially matched the prediction confidence for outliers with the confidence for inliers.

OSLO achieves the best trade-off. Benchmark results in Table 4.1 show that OSLO surpasses the best transductive methods in terms of closed-set accuracy, while consistently outperforming existing out-of-distribution and open-set detection competitors on outlier detection ability. Interestingly, while the gap between closed-set accuracy of transductive methods and inductive ones typically contracts with more shots, the outlier detection performance of OSLO remains largely superior to its inductive competitors even in the 5-shot scenario, where a consistent 3-6% gap in AUROC and AUPR with the second-best method can be observed. We accumulate further evidence of OSLO's superiority by introducing 3 additional cross-domain scenarios in Figure 4.2, corresponding to a base model pre-trained on *tiered*-ImageNet, but tested on

Table 4.2 **OSLO's ablation study** along two factors described in Table 4.5.2. Results are produced on the 1-shot scenario on *mini*-ImageNet, with a ResNet-12

(i) *Inlierness* latent	Acc	AUROC
Ignore (4.2)	69.42	64.97
Leverage (4.4)	71.73	74.92
(ii) Optimization steps	Acc	AUROC
At initialization	66.63	71.76
After optimization	71.73	74.92

CUB, Aircraft, and Fungi datasets. In such challenging scenarios, where both feature and class distributions shift, OSLO remains competitive in closed-set accuracy and largely outperforms other methods in outlier detection.

OSLO benefits from more query samples. A critical question for transductive methods is the dependency of their performance on the size of the query set. Intuitively, a larger query set will provide more unlabeled data and thus lead to better results. We exhibit this relation in Figure 4.3 by spanning the number of queries per class from 1 to 30. We observe that the closed-set accuracy of most transductive methods is stable across this span in the 5-shot scenario. In the 1-shot scenario, OSLO gains from additional queries but stays above the baseline even with a small number of queries. Interestingly enough, OSLO is the only transductive method to improve its outlier detection ability when the number of queries increases.

OSLO steps toward model-agnosticity. We evaluate OSLO's *model-agnosticity* by its ability to maintain consistent improvement over the *Strong Baseline*, regardless of the model used, and without hyperparameter adjustment. In that regard, we depart from the standard ResNet-12 and cover 3 largely distinct architectures, each encoding different inductive biases. To further strengthen our empirical demonstration of OSLO's model-agnosticity, for each architecture, we consider several training strategies spanning different paradigms – unsupervised, supervised, semi and semi-weakly supervised – and using different types of data –image, text–. Results in Figure 4.4 show the relative improvement of OSLO w.r.t the strong baseline in the 1-shot

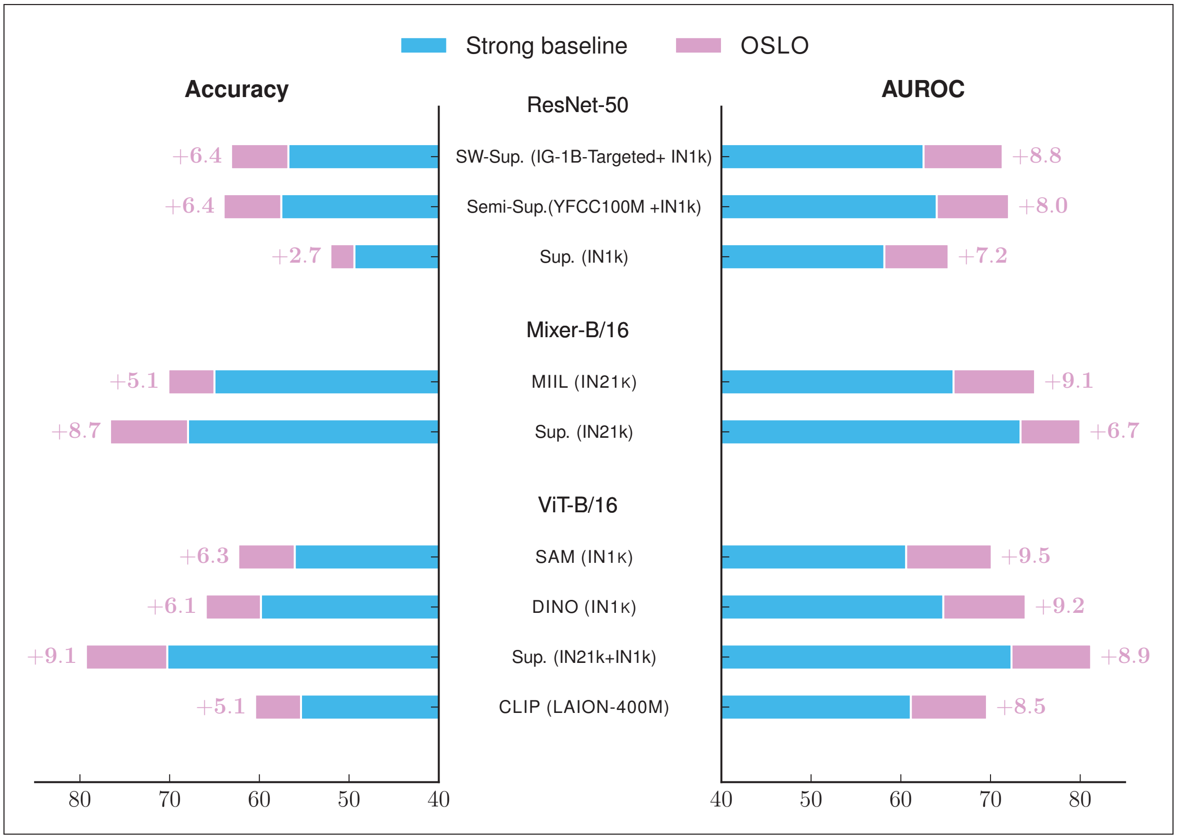

Figure 4.4 **OSLO's improvement is consistent across many architectures and training strategies.** To evaluate model-agnosticity, we compare OSLO to the Strong baseline on challenging 1-shot Fungi tasks. We experiment across 3 largely distinct architectures: ResNet-50 (CNN) (He *et al.*, 2016), ViT-B/16 (Vision Transformer) (Dosovitskiy *et al.*, 2021), and Mixer-B/16 (MLP-Mixer) (Tolstikhin *et al.*, 2021). For each architecture, we include different types of pre-training, including Supervised (Sup.), Semi-Supervised, Semi-Weakly Supervised (SW Sup.) (Yalniz *et al.*, 2019), DINO (Caron *et al.*, 2021), SAM (Chen *et al.*, 2022), MIIL (Ridnik *et al.*, 2021). Improvements over the baseline are consistently significant and generally higher than those observed with the ResNet-12 in Figure 4.2

scenario on the $*$ → Fungi benchmark. Without any tuning, OSLO remains able to leverage the strong expressive power of large-scale models, and even consistently widens the gap with the strong baseline, achieving a remarkable performance of 79% accuracy and 81% AUROC with the ViT-B/16. This set of results testifies to how easy obtaining highly competitive results on difficult specialized tasks can be by combining OSLO with the latest models.

Ablation study. We perform an ablation study on the important ingredients of OSLO. As a core contribution of our work, we show in Table 4.2 that the presence *and* optimization of the latent *inlierness* scores is crucial. In particular, the closed-form latent score ξ yields strong outlier recognition performance, even at *initialization* (i.e. after the very first update from Prop. 4.4.0.1). Interestingly, refining the parametric model without accounting for $\boldsymbol{\xi}$ in $\mathbf{Z}$ and $\boldsymbol{\mu}$'s updates (i.e. standard likelihood) allows the model to fit those outliers, leading to significantly worse outlier detection, from 71.76% to 64.97%. On the other hand, accounting for $\boldsymbol{\xi}$, as proposed in OSLO, improves the outlier detection by more than 3% over the initial state, and closed-set accuracy by more than 5%. In the end, in a fully apples-to-apples comparison, OSLO outperforms its standard likelihood counterpart by more than 2% in accuracy and 10% in outlier detection.

4.6 Discussion

Limitations. Unlike inductive methods, transductive methods are inevitably affected by the amount of unlabelled data provided, which in real-world scenarios cannot necessarily be controlled. OSLO is no exception, and fewer query samples tend to decrease its performance. In the extreme case with only 1 sample per class, OSLO's performance comes close to our inductive baseline. In those scenarios where unlabelled data is particularly scarce, the benefits brought by transduction remain therefore limited. As a second limitation, poorer representations appear to diminish OSLO's competitive advantage in closed-set accuracy. In particular, OSLO's closed-set accuracy stands more than 6% above the baseline in the in-domain *tiered* → *tiered* scenario but reduces to 2% in the most challenging-domain scenario *tiered* → *Aircraft*. Figure 4.4 further corroborates this hypothesis, with OSLO's accuracy outperforming the baseline's by 9% with the best transformer, but by only 2.7% on the least performing model. Finally, although not within the scope of the current paper, the presence of class imbalance within the query sets should further be considered to obtain a truly realistic setting. In that case, the class-balance assumption made in section 4.4 would need to be adapted with techniques from Chapters 3 and 5.

Conclusion. We presented OSLO, the first transductive method for FSOSR. OSLO extends the vanilla maximum likelihood objective in two important ways, First, it accounts for the constraints imposed by the provided supervision. More importantly, it explicitly models the potential presence of outliers in its very latent model, allowing it to co-learn the optimal closed-set model and outlier assignments. Beyond FSOSR, we believe OSLO presents a general, conceptually simple, and completely modular formulation to leverage unlabelled data in the potential presence of outliers. That, of course, naturally extends to other classification settings, such as large-scale open-set detection, but to other tasks as well, such as segmentation in which *background* pixels could be viewed as *outliers* with respect to closed-set classes. We hope OSLO inspires further work in that direction.

CHAPTER 5

FEW-SHOT SEGMENTATION WITHOUT META-LEARNING: A GOOD TRANSDUCTIVE INFERENCE IS ALL YOU NEED?

Malik Boudiaf[1] , Ismail Ben Ayed[1] , Hoel Kervadec[1] , Imtiaz Masud Ziko[1] , Pablo Piantanida[2] , Jose Dolz[1]

[1] ÉTS Montréal, QC, Canada,
[2] Laboratoire des Signaux et Systèmes (L2S),
CentraleSupelec-CNRS-Université Paris-Saclay, France

Poster at Conference on Computer Vision and Pattern Recognition (CVPR) 2021, Nashville.

Abstract

Few-shot segmentation has recently attracted substantial interest, with the popular meta-learning paradigm widely dominating the literature. We show that the way inference is performed for a given few-shot segmentation task has a substantial effect on performances, an aspect that has been overlooked in the literature. We introduce a transductive inference, which leverages the statistics of the unlabeled pixels of a task by optimizing a new loss containing three complementary terms: (i) a standard cross-entropy on the labeled pixels; (ii) the entropy of posteriors on the unlabeled query pixels; and (iii) a global KL-divergence regularizer based on the proportion of the predicted foreground region. Our inference uses a simple linear classifier of the extracted features, has a computational load comparable to inductive inference and can be used on top of any base training. Using standard cross-entropy training on the base classes, our inference yields highly competitive performances on well-known few-shot segmentation benchmarks[13]. On PASCAL-5^i, it brings about 5% improvement over the best performing state-of-the-art method in the 5-shot scenario, while being on par in the 1-shot setting. Even more surprisingly, this gap widens as the number of support samples increases, reaching up to 6% in the 10-shot scenario. Furthermore, we introduce a more realistic setting with domain shift, where the base and novel classes are drawn from different data sets. In this setting, we found that our method achieves the best performances.

[13] The code is provided with this submission.

5.1 Introduction

Few-shot learning, which aims at classifying instances from unseen classes given only a handful of training examples, has witnessed a rapid progress in the recent years. In order to quickly adapt to novel classes, there has been a substantial focus on the meta-learning (or learning-to-learn) paradigm (Ren *et al.*, 2018; Snell *et al.*, 2017; Vinyals *et al.*, 2016b). Meta-learning approaches have popularized the need of structuring the training data into *episodes*, thereby simulating the tasks that will be presented at inference. Nevertheless, despite the achieved improvements, several recent image classification works (Boudiaf *et al.*, 2020a; Chen *et al.*, 2019; Dhillon *et al.*, 2020; Guo *et al.*, 2020; Tian *et al.*, 2020a; Ziko *et al.*, 2020) observed that meta-learning approaches might have limited generalization capacity beyond the standard 1-shot or 5-shot classification benchmarks. For instance, in more realistic setting with domain-shift, simple classification baselines may outperform much more complex meta-learning methods (Chen *et al.*, 2019; Guo *et al.*, 2020).

Deep-learning based semantic segmentation has been generally nurtured from the methodological advances in image classification. Few-shot segmentation, which has gained considerable popularity in the recent years (Gairola, Hemani, Chopra & Krishnamurthy, 2020; Li, Wei, Chen, Tai & Tang, 2020; Liu *et al.*, 2020d; Nguyen & Todorovic, 2019; Rakelly *et al.*, 2018; Tian *et al.*, 2020b; Wang *et al.*, 2020a, 2019a; Yang *et al.*, 2020a; Yang *et al.*, 2020d; Zhang, Lin, Liu, Yao & Shen, 2019b; Zhang *et al.*, 2020b), is no exception. In this setting, a deep segmentation model is first pre-trained on *base* classes. Then, model generalization is assessed over few-shot *tasks* and novel classes unseen during base training. Each task includes an unlabeled test image, referred to as the *query*, along with a few labeled images (the *support* set). The recent literature in few-shot segmentation follows the learning-to-learn paradigm, and substantial research efforts focused on the design of specialized architectures and episodic-training schemes for base training. However, (i) episodic training itself implicitly assumes that testing tasks have a similar structure (e.g., the number of support shots) to the tasks used at the meta-training stage, and (ii) both base and novel classes are often assumed to be sampled from the same dataset.

In practice, those assumptions may limit the applicability of the existing few-shot segmentation methods in realistic scenarios (Cao *et al.*, 2020; Chen *et al.*, 2019). In fact, our experiments proved consistent with findings in few-shot classification when going beyond the standard settings and benchmarks. Particularly, we observed among state-of-the-art methods a saturation in performances (Cao *et al.*, 2020) when increasing the number of labeled samples (See Table 5.3). Also, in line with very recent observations in image classification (Chen *et al.*, 2019), existing meta-learning methods prove less competitive in cross-domains scenarios (See Table 5.4).

This casts doubts as to the viability of the current few-shot segmentation benchmarks and datasets, and suggests that the recent progress in performances might be, to a large extent, due to carefully designed architectures and episodic-training schemes. Also, this motivates re-considering the existing benchmarks and re-thinking the relevance of the meta-learning paradigm, which has become the *de facto* choice in the few-shot segmentation literature.

In this work, we forego meta-learning, and re-consider a simple cross-entropy supervision during training on the base classes for feature extraction. We show that the way inference is performed has a substantial effect on performances, an aspect that, to our knowledge, has been completely overlooked in the few-shot segmentation literature. Specifically, we propose a new *transductive* inference based on a linear classifier built on top of the extracted features. Unlike *inductive* inference, our transductive setting also exploits the unlabeled pixels from the query image, which we naturally have access to, when building the classifier for a task. Therefore, our inference leverages the structure and global statistics of both the unlabeled and labeled pixels of a given few-shot segmentation task by optimizing an original task-specific loss function. We emphasize that we have access to exactly the same information as in standard inductive inference for a given few-shot segmentation task (i.e., one query image and a few labeled support images), and do not use any additional unlabeled data. Fig. 5.1 depicts the results of a standard inductive inference using support-based initial classifier weights (third column) or additional cross-entropy fine-tuning on the support set (fourth column), and juxtapose them to the much enhanced results obtained with our transductive inference (last column).

Contributions

- We present a new transductive inference for a given few-shot segmentation task, which optimizes a loss integrating three complementary terms: *i)* a standard cross-entropy on the labeled pixels of the support images; *ii)* the entropy of the posteriors on the query pixels of the test image; and *iii)* a global KL divergence regularizer based on the proportion of the predicted foreground pixels within the test image. Our transductive inference is based on a simple linear classifier of the extracted features, has a computational load comparable to inductive inference and is modular: it can be used on top of any trained feature extractor. We call our inference RePRI (Region Proportion Regularized Inference).
- Although we use a basic cross-entropy training on the base classes, without complex meta-learning schemes, RePRI yields highly competitive performances on the standard few-shot segmentation benchmarks, PASCAL-5^i and COCO-20^i. Particularly, on PASCAL-5^i, we report a gain of almost 5% with respect to the state-of-the-art in the 5-shot scenario, while being on par with it in the 1-shot setting. This gap consistently widens as the number of support samples increases, reaching up to 6% in the 10-shot scenario. This suggests that our transductive inference leverages more effectively the information from the labeled support set of a task.
- We introduce a more realistic setting where, in addition to the usual shift on classes between training and testing data distributions, a shift on the images' feature distribution is also introduced. We find that our method achieves the best performances in this scenario.
- We demonstrate that a precise region-proportion information on the query object improves substantially the results, with an average gain of 13% on both datasets. While assuming the availability of such information is not realistic, we show that inexact estimates can still lead to drastic improvements, opening a very promising direction for future research.

5.2 Related Work

5.2.1 Few shot Learning for classification

Meta-learning has become the *de facto* solution to learn novel tasks from a few labeled samples. Even though the idea is not new (Schmidhuber, 1987), it has been revived recently by several popular works in few-shot classification (Finn *et al.*, 2017; Ravi & Larochelle, 2016; Ren *et al.*, 2018; Snell *et al.*, 2017; Vinyals *et al.*, 2016b). These works can be categorized into gradient- or metric-learning-based methods. On the one hand, gradient approaches resort to stochastic gradient descent (SGD) to learn the commonalities among different tasks (Ravi & Larochelle, 2016; Finn *et al.*, 2017). On the other hand, metric-learning approaches (Vinyals *et al.*, 2016b; Snell *et al.*, 2017) adopt deep networks as feature-embedding functions, and compare the distances between the embeddings. Furthermore, in a recent line of works, the transductive setting has been investigated for few-shot classification (Dhillon *et al.*, 2020; Boudiaf *et al.*, 2020a; Hou *et al.*, 2019; Kim *et al.*, 2019; Liu *et al.*, 2019c; Qiao *et al.*, 2019; Snell *et al.*, 2017; Ziko *et al.*, 2020), and yielded performance improvements over inductive inference. These results are in line with established facts in classical transductive inference (Vapnik, 1999; Joachims, 1999; Dengyong *et al.*, 2004), well-known to outperform its inductive counterpart on small training sets. To a large extent, these transductive classification works follow well-known concepts in semi-supervised learning, such as graph-based label propagation (Liu *et al.*, 2019c), entropy minimization (Dhillon *et al.*, 2020) or Laplacian regularization (Ziko *et al.*, 2020). While the entropy is a part of our transductive loss, we show that it is not sufficient for segmentation tasks, typically yielding trivial solutions.

5.2.2 Few-shot segmentation

Segmentation can be viewed as a classification at the pixel level, and recent efforts mostly went into the design of specialized architectures for few-shot segmentation. Typically, the existing methods use a two-branch comparison framework, inspired from the very popular prototypical networks for few-shot classification (Snell *et al.*, 2017). Particularly, the support images are employed to generate class prototypes, which are later used to segment the query images via a prototype-query comparison module. Early frameworks followed a dual-branch architecture, with two independent branches (Shaban *et al.*, 2018; Dong & Xing, 2018; Rakelly *et al.*, 2018), one generating the prototypes from the support images and the other segmenting the query images with the learned prototypes. More recently, the dual-branch setting has been unified into a single-branch architecture, which employs the same embedding function for both the support and query sets (Zhang *et al.*, 2020b; Siam *et al.*, 2019; Wang *et al.*, 2019a; Yang *et al.*, 2020a; Liu *et al.*, 2020f). These approaches mainly aim at exploiting better guidance for the segmentation of query images (Zhang *et al.*, 2020b; Nguyen & Todorovic, 2019; Wang *et al.*, 2020a; Zhang *et al.*, 2019a), by learning better class-specific representations (Wang *et al.*, 2019a; Liu *et al.*, 2020d,f; Yang *et al.*, 2020a; Siam *et al.*, 2019) or iteratively refining these (Zhang *et al.*, 2019b). Graph CNNs have also been employed to establish more robust correspondences between the support and query images, enhancing the learned prototypes (Wang *et al.*, 2020a). Alternative solutions to learn better class representations include: imprinting the weights for novel classes (Siam *et al.*, 2019), decomposing the holistic class representation into a set of part-aware prototypes (Liu *et al.*, 2020f) or mixing several prototypes, each corresponding to diverse image regions (Yang *et al.*, 2020a).

5.3 Formulation

5.3.1 Few-shot Setting

Formally, we define a *base* dataset $\mathcal{D}_{\text{base}}$ with base semantic classes $\mathcal{Y}_{\text{base}}$, employed for training. Specifically, $\mathcal{D}_{\text{base}} = \{(x_n, y_n)\}_{n=1}^{N}$, $\Omega \subset \mathbb{R}^2$ an image space, $x_n : \Omega \rightarrow \mathbb{R}^3$ an input image, and $y_n : \Omega \rightarrow \{0, 1\}^{|\mathcal{Y}_{base}|}$ its corresponding pixelwise one-hot annotation. At inference, we test our model through a series of K-shots tasks. Each K-shots task consists of a *support* set $\mathbb{S} = \{(x_k, y_k)\}_{k=1}^{K}$, i.e. K fully annotated images, and one unlabeled query image $x_{\mathbb{Q}}$, all from the same novel class. This class is randomly sampled from a set of *novel* classes $\mathcal{Y}_{\text{novel}}$ such that

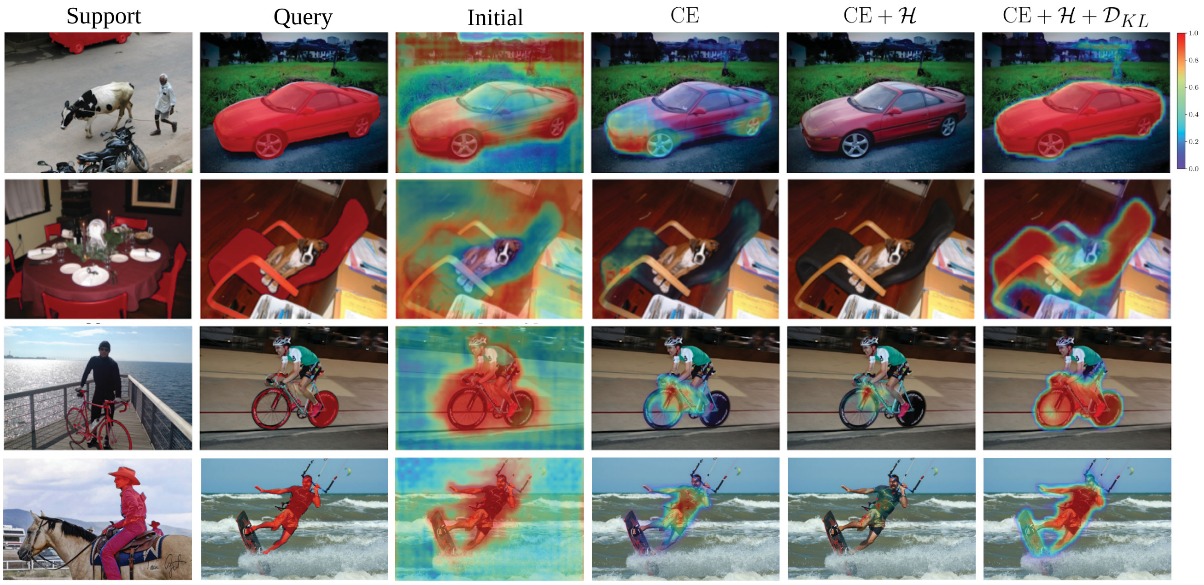

Figure 5.1 Probability maps for several 1-shot tasks. For each task, the two first columns show the ground truth of support and query. *Initial* column represents the probability map with the initial classifier $\boldsymbol{\theta}^{(0)}$, and the last three columns show the final soft predicted segmentation after finetuning with each of the three losses. Best viewed in colors

$\mathcal{Y}_{\text{base}} \cap \mathcal{Y}_{\text{novel}} = \emptyset$. The goal is to leverage the supervision provided by the support set in order to properly segment the object of interest in the query image.

5.3.2 Base training

5.3.3 Inductive bias in episodic training

There exist different ways of leveraging the base set $\mathcal{D}_{\text{base}}$. Meta-learning, or *learning to learn*, is the dominant paradigm in the few-shot literature. It emulates the test-time scenario during training by structuring $\mathcal{D}_{\text{base}}$ into a series of training tasks. Then, the model is trained on these tasks to learn how to best leverage the supervision from the support set in order to enhance its query segmentation. Recently, Cao et al. (Cao *et al.*, 2020) formally proved that the number of shots K_{train} used in training episodes in the case of prototypical networks represents a learning bias, and that the testing performance saturates quickly when K_{test} differs from K_{train}.

Empirically, we observed the same trend for current few-shot segmentation methods, with minor improvements from 1-shot to 5-shot performances (Table 5.1).

5.3.4 Standard training

In practice, the format of the test tasks may be unknown beforehand. Therefore, we want to take as few assumptions as possible on this. This motivates us to employ a feature extractor $f_{\boldsymbol{\phi}}$ trained with standard cross-entropy supervision on the whole $\mathcal{D}_{\text{base}}$ set instead, **without resorting to episodic training.**

5.3.5 Inference

Objective: In what follows, we use $\bullet$ as a placeholder for either k or Q. At inference, we consider the 1-way segmentation problem: $y_{\bullet}$ represents the foreground/background (F/B) of image $x_{\bullet}$, i.e. $y_{\bullet} : \Omega \rightarrow \{0, 1\}^2$. For both support and query images, we extract features $z_{\bullet} := f_{\boldsymbol{\phi}}(x_{\bullet})$ and $z_{\bullet} : \Psi \rightarrow \mathbb{R}^C$, where C is the channel dimension in the feature space Ψ, with lower resolution $|\Psi| < |\Omega|$. We also introduce the down-sampling operator $\widetilde{\bullet} : \Omega \rightarrow \Psi$.

Using features $z_{\bullet}$, our goal is to build a classifier $\boldsymbol{\theta}$ that properly discriminates foreground from background pixels. Precisely, for every pixel location j, we want to model the probability $p_{\bullet}(j) := \mathbb{P}\left(\widetilde{y}_{\bullet}(j) \left| z_{\bullet(j)}; \boldsymbol{\theta}\right.\right)$ parametrized by learnable parameters $\boldsymbol{\theta}$. Notice that $p_{\bullet} : \Psi \rightarrow [0, 1]^2$. To obtain a final segmentation for the query image $x_{\mathbb{Q}}$, we resort to an up-sampling operator $\acute{\bullet} : \Psi \rightarrow \Omega$. For metrics computation, we compare $\acute{p}_{\mathbb{Q}}$ to $y_{\mathbb{Q}}$ (note that the query label is only used at evaluation).

In order to find the weights of the classifier, we propose to optimize the following transductive-inference objective:

$$\min_{\boldsymbol{\theta}} \ \mathrm{CE} + \lambda(\mathcal{H} + \mathcal{D}_{KL}), \tag{5.1}$$

where $\lambda \in \mathbb{R}$ is a non-negative hyper-parameter balancing the effects of the different terms. The following describes in detail each of the terms in (5.1).

- CE $= -\frac{1}{K|\Psi|}\sum_{k=1}^{K}\sum_{j\in\Psi}\widetilde{y}_k(j)\cdot\log(p_k(j))$ is the cross-entropy between the pixel labels from support images in $\mathbb{S}$ and their corresponding softmax predictions, and $\cdot : [0,1]^2 \times [0,1]^2 \rightarrow \mathbb{R}$ the dot product operation. Simply minimizing this term will often lead to degenerate solutions, especially in the 1-shot setting, as observed in Figure 5.1 – the classifier $\boldsymbol{\theta}$ typically overfits the support set $\mathbb{S}$, translating into small activated regions on the query image.

- $\mathcal{H} = -\frac{1}{|\Psi|}\sum_{j\in\Psi} p_{\mathbb{Q}}(j)\cdot\log\left(p_{\mathbb{Q}}(j)\right)$ is the entropy of the predictions on the query pixels. The role of this entropy term is to make the model's predictions more confident on the query image. The use of $\mathcal{H}$ originates from the semi-supervised literature (Grandvalet & Bengio, 2005; Miyato *et al.*, 2018; Berthelot *et al.*, 2019). Intuitively, it pushes the decision boundary drawn by the linear classifier toward low-density regions of the extracted query feature space. While this term plays a crucial role in conserving object regions that were initially predicted with only medium confidence, its sole addition to CE does not solve the problem of degenerate solutions, and may even worsen it in some cases.

- $\mathcal{D}_{\text{KL}} = \widehat{p}_{\mathbb{Q}}\cdot\log\left(\frac{\widehat{p}_{\mathbb{Q}}}{\boldsymbol{\pi}}\right)$, with $\widehat{p}_{\mathbb{Q}} = \frac{1}{|\Psi|}\sum_{j\in\Psi} p_{\mathbb{Q}}(j)$, is a Kullback-Leibler (KL) Divergence term that encourages the F/B proportion predicted by the model to match a parameter $\boldsymbol{\pi} \in [0,1]^2$. The joint estimation of parameter $\boldsymbol{\pi}$ in our context is further discussed in the following subsection. Here, we argue that this term plays a **key role** in our loss. First, in the case where parameter $\boldsymbol{\pi}$ does not match the exact F/B proportion of the query image, this term still helps avoid the degenerate solutions stemming from CE and $\mathcal{H}$ minimization. And second, should an accurate estimate of the F/B proportion in the query image be available, it could easily be embedded through this term, resulting in a substantial performance boost, as discussed in Section 5.4.

Choice of the classifier: As we optimize $\boldsymbol{\theta}$ for each task at inference, we want our method to add as little computational load as possible. In this regard, we employ a simple linear classifier with learnable parameters $\boldsymbol{\theta}^{(t)} = \{\boldsymbol{w}^{(t)}, b^{(t)}\}$, with t the current step of the optimization procedure and where $\boldsymbol{w}^{(t)} \in \mathbb{R}^C$ represents the *foreground* prototype and $b^{(t)} \in \mathbb{R}$ the corresponding bias.

Thus, the probabilities $p^{(t)}$ for iteration t can be obtained as follow:

$$p^{(t)}_{\boldsymbol{\cdot}}(j) := \begin{pmatrix} 1 - s^{(t)}_{\boldsymbol{\cdot}}(j) \\ s^{(t)}_{\boldsymbol{\cdot}}(j) \end{pmatrix}, \tag{5.2}$$

where $s^{(t)}_{\boldsymbol{\cdot}}(j) = \text{sigmoid}\left(\tau\left[\cos\left(z_{\boldsymbol{\cdot}}(j), \boldsymbol{w}^{(t)}\right) - b^{(t)}\right]\right)$, $\tau \in \mathbb{R}$ is a temperature hyper-parameter, and cos the cosine similarity. The same classifier is used to estimate the support set probabilities p_k and the query predicted probabilities $p_{\mathbb{Q}}$. At initialization, we set prototype $\boldsymbol{w}^{(0)}$ to be the average of the foreground support features, i.e. $\boldsymbol{w}^{(0)} = \frac{1}{K|\Psi|}\sum_{k=1}^{K}\sum_{j\in\Psi}\widetilde{y}_k(j)_1 z_k(j)$. Initial bias $b^{(0)}$ is set as the mean of the foreground's soft predictions on the query image: $b^{(0)} = \frac{1}{|\Psi|}\sum_{j\in\Psi} p_{\mathbb{Q}}(j)_1$. Then, $\boldsymbol{w}^{(t)}$ and $b^{(t)}$ are optimized with gradient descent, whose computational footprint is discussed in Section 5.4.

Joint estimation of F/B proportion $\boldsymbol{\pi}$: Without additional information, we leverage the model's label-marginal distribution over the query image $\widehat{p}^{(t)}_{\mathbb{Q}}$ in order to learn $\boldsymbol{\pi}$ jointly with classifier parameters. Note that minimizing Eq. 5.1 with respect to $\boldsymbol{\pi}$ yields $\boldsymbol{\pi}^{(t)} = \widehat{p}^{(t)}_{\mathbb{Q}}$. Empirically, we found that after initialization, updating $\boldsymbol{\pi}$ only once during optimization, at a later iteration $t_{\boldsymbol{\pi}}$ was enough, such that:

$$\boldsymbol{\pi}^{(t)} = \begin{cases} \widehat{p}^{(0)}_{\mathbb{Q}} & 0 \leq t \leq t_{\boldsymbol{\pi}} \\ \widehat{p}^{(t_{\boldsymbol{\pi}})}_{\mathbb{Q}} & t > t_{\boldsymbol{\pi}}. \end{cases} \tag{5.3}$$

Intuitively, the entropy term $\mathcal{H}$ helps gradually refine initially blurry soft predictions (third column in Fig. 5.1), which turns $\widehat{p}^{(t)}_{\mathbb{Q}}$ into an improving estimate of the true F/B proportion. A quantitative study of this phenomenon is provided in Section 5.4.5. Therefore, our inference can be seen as a joint optimization over $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$, with $\mathcal{D}_{\text{KL}}$ serving as a *self-regularization* that prevents the model's marginal distribution $\widehat{p}^{(t)}_{\mathbb{Q}}$ from diverging.

Oracle case with a known $\boldsymbol{\pi}$: As an upper bound, we also investigate the *oracle* case, where we have access to the true F/B proportion in $x_{\mathbb{Q}}$:

$$\boldsymbol{\pi}^* = \frac{1}{|\Psi|} \sum_{j \in \Psi} \widetilde{y}_{\mathbb{Q}}(j). \tag{5.4}$$

5.4 Experiments

5.4.1 Experimental setup

Datasets. We resort to two public few-shot segmentation benchmarks, PASCAL-5^i and COCO-20^i, to evaluate our method. PASCAL-5^i is built from PascalVOC 2012 (Everingham, Van Gool, Williams, Winn & Zisserman, 2010), and contains 20 object categories split into 4 folds. For each fold, 15 classes are used for training and the remaining 5 categories for testing. COCO-20^i is built from MS-COCO (Lin *et al.*, 2014) and is more challenging, as it contains more samples, more classes and more instances per image. Similar to PASCAL-5^i, COCO-20^i dataset is divided into 4-folds with 60 base classes and 20 test classes in each fold.

Training. We build our model based on PSPNet (Zhao, Shi, Qi, Wang & Jia, 2017) with Resnet-50 and Resnet-101 (He *et al.*, 2016) as backbones. We train the feature extractor with standard cross-entropy over the base classes during 100 epochs on PASCAL-5^i, and 20 epochs on COCO-20^i, with batch size set to 12. We use SGD as optimizer with the initial learning rate set to 2.5e−3 and we use cosine decay. Momentum is set to 0.9, and weight decay to 1e−4. Label smoothing is used with smoothing parameter $\epsilon = 0.1$. We did not use multi-scaling, nor deep supervision, unlike the original PSPNet paper (Zhao *et al.*, 2017). As for data augmentations, we only use random mirror flipping.

Inference. At inference, following previous works (Liu *et al.*, 2020f; Wang *et al.*, 2019a), all images are resized to a fixed 417×417 resolution. For each task, the classifier $\boldsymbol{\theta}$ is built on top of the features from the penultimate layer of the trained network. For our model with ResNet-50

Table 5.1 Results of 1-way 1-shot and 1-way 5-shot segmentation on PASCAL-5^i using the mean-IoU. Best results in bold

		1 shot					5 shot				
Method	Backbone	Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	Mean
OSLSM (Shaban *et al.*, 2018)	VGG-16	33.6	55.3	40.9	33.5	40.8	35.9	58.1	42.7	39.1	43.9
co-FCN (Rakelly *et al.*, 2018)		36.7	50.6	44.9	32.4	41.1	37.5	50.0	44.1	33.9	41.4
AMP (Siam *et al.*, 2019)		41.9	50.2	46.7	34.7	43.4	41.8	55.5	50.3	39.9	46.9
PANet (Wang *et al.*, 2019a)		42.3	58.0	51.1	41.2	48.1	51.8	64.6	59.8	46.5	55.7
FWB (Nguyen & Todorovic, 2019)		47.0	59.6	52.6	48.3	51.9	50.9	62.9	56.5	50.1	55.1
SG-One (Zhang *et al.*, 2020b)		40.2	58.4	48.4	38.4	46.3	41.9	58.6	48.6	39.4	47.1
CRNet (Liu *et al.*, 2020d)		-	-	-	-	55.2	-	-	-	-	58.5
FSS-1000 (Li *et al.*, 2020)		-	-	-	-	-	37.4	60.9	46.6	42.2	56.8
RPMM (Liu *et al.*, 2020f)		47.1	65.8	50.6	48.5	53.0	50.0	66.5	51.9	47.6	54.0
CANet (Zhang *et al.*, 2019b)	ResNet50	52.5	65.9	51.3	51.9	55.4	55.5	67.8	51.9	53.2	57.1
PGNet (Zhang *et al.*, 2019a)		56.0	66.9	50.6	50.4	56.0	57.7	68.7	52.9	54.6	58.5
CRNet (Liu *et al.*, 2020d)		-	-	-	-	55.7	-	-	-	-	58.8
SimPropNet (Gairola *et al.*, 2020)		54.9	67.3	54.5	52.0	57.2	57.2	68.5	58.4	56.1	60.0
LTM (Yang *et al.*, 2020d)		52.8	69.6	53.2	52.3	57.0	57.9	69.9	56.9	57.5	60.6
RPMM (Yang *et al.*, 2020a)		55.2	66.9	52.6	50.7	56.3	56.3	67.3	54.5	51.0	57.3
PPNet (Liu *et al.*, 2020f)*		47.8	58.8	53.8	45.6	51.5	58.4	67.8	64.9	56.7	62.0
PFENet (Tian *et al.*, 2020b)		**61.7**	**69.5**	55.4	**56.3**	**60.8**	63.1	70.7	55.8	57.9	61.9
RePRI (ours)		59.8	68.3	**62.1**	48.5	59.7	**64.6**	**71.4**	**71.1**	**59.3**	**66.6**
Oracle-RePRI	ResNet50	72.4	78.0	77.1	65.8	73.3	75.1	80.8	81.4	74.4	77.9
FWB (Nguyen & Todorovic, 2019)	ResNet101	51.3	64.5	56.7	52.2	56.2	54.9	67.4	62.2	55.3	59.9
DAN (Wang *et al.*, 2020a)		54.7	68.6	57.8	51.6	58.2	57.9	69.0	60.1	54.9	60.5
PFENet (Tian *et al.*, 2020b)		**60.5**	**69.4**	54.4	**55.9**	**60.1**	62.8	70.4	54.9	57.6	61.4
RePRI (ours)		59.6	68.6	**62.2**	47.2	59.4	**66.2**	**71.4**	**67.0**	**57.7**	**65.6**

* We report the results where no additional unlabeled data is employed.

as backbone, this results in a $53 \times 53 \times 512$ features map. SGD optimizer is used to train $\boldsymbol{\theta}$, with a learning rate of 0.025. For each task, a total of 50 iterations are performed. The parameter update iteration $t_{\boldsymbol{\pi}}$ is set to 10. The weight λ is automatically set to $1/K$, such that the CE term plays a more important role as the number of shots K grows. Finally, the temperature τ is set to 20.

Evaluation. We employ the widely adopted mean Intersection over Union (mIoU). Specifically, for each class, the classwise-IoU is computed as the sum over all samples within the class of the intersection over the sum of all unions. Then, the mIoU is computed as the average over all classes of the classwise-IoU. Following previous works (Liu *et al.*, 2020f), 5 runs of 1000 tasks each are computed for each fold, and the average mIoU over runs is reported.

Table 5.2 Results of 1-way 1-shot and 1-way 5-shot segmentation on COCO-20^i using mean-IoU metric. Best results in bold

		1 shot					5 shot				
Method	Backbone	Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	Mean
PPNet* (Liu *et al.*, 2020f)		34.5	25.4	24.3	18.6	25.7	48.3	30.9	35.7	30.2	36.2
RPMM (Yang *et al.*, 2020a)	ResNet50	29.5	36.8	29.0	27.0	30.6	33.8	42.0	33.0	33.3	35.5
PFENet (Tian *et al.*, 2020b)		**36.5**	38.6	**34.5**	**33.8**	**35.8**	36.5	43.3	37.8	38.4	39.0
RePRI (ours)		32.0	**38.7**	32.7	33.1	34.1	**39.3**	**45.4**	**39.7**	**41.8**	**41.6**
Oracle-RePRI	ResNet50	49.3	51.4	38.2	41.6	45.1	51.5	60.8	54.7	55.2	55.5

* We report the results where no additional unlabeled data is employed.

5.4.2 Benchmark results

Main method. First, we investigate the performance of the proposed method in the popular 1-shot and 5-shot settings on both PASCAL-5^i and COCO-20^i, whose results are reported in Table 5.1 and 5.2. Overall, we found that our method compares competitively with state-of-the-art approaches in the 1-shot setting, and significantly outperforms recent methods in the 5-shot scenario. Additional qualitative results on PASCAL-5^i are shown in Fig. 5.3.

Beyond 5-shots. In the popular learning-to-learn paradigm, the number of shots leveraged during the meta-training stage has a direct impact on the performance at inference (Cao *et al.*, 2020). Particularly, to achieve the best performance, meta-learning based methods typically require the numbers of shots used during meta-training to match those employed during meta-testing. To demonstrate that the proposed method is more robust against differences on the number of labeled support samples between the base and test sets, we further investigate the 10-shot scenario. Particularly, we trained the methods in (Tian *et al.*, 2020b; Yang *et al.*, 2020a) by using one labeled sample per class, i.e., 1-shot task, and test the models on a 10-shots task. Interestingly, we show that the gap between our method and current state-of-the-art becomes larger as the number of support images increases (Table 5.3), with significant gains of 6% and 4% on PASCAL-5^i and COCO-20^i, respectively. These results suggest that our transductive inference leverages more effectively the information conveyed in the labeled support set of a given task.

Table 5.3 Aggregated results for 1-way 1-, 5- and 10-shot tasks with Resnet50 as backbone and averaged over 4 folds. Per fold results are available in the supplementary material

	PASCAL-5^i			COCO-20^i		
Method	1-S	5-S	10-S	1-S	5-S	10-S
RPMM (Yang *et al.*, 2020a)	56.3	57.3	57.6	30.6	35.5	33.1
PFENet (Tian *et al.*, 2020b)	**60.8**	61.9	62.1	**35.8**	39.0	39.7
RePRI (ours)	59.7	**66.6**	**68.1**	34.1	**41.6**	**44.1**
Oracle-RePRI	73.3	77.9	78.6	45.1	55.5	58.7

5.4.3 Oracle results

We now investigate the ideal scenario where an oracle provides the exact foreground/background proportion in the query image, such that $\boldsymbol{\pi}^{(t)} = \boldsymbol{\pi}^*, \forall t$. Reported results in this scenario, referred to as *Oracle* (Table 5.1 and 5.2) show impressive improvements over both our current method and all previous works, with a consistent gain across datasets and tasks. Particularly, these values range from 11% and 14 % on both PASCAL-5^i and COCO-20^i and in both 1-shot and 5-shot settings. We believe that these findings convey two important messages. First, it proves that there exists a simple linear classifier that can largely outperform state-of-the-art meta-learning models, while being built on top of a feature extractor trained with a standard cross-entropy loss. Second, these results indicate that having a precise size of the query object of interest acts as a strong regularizer. This suggests that more efforts could be directed towards properly constraining the optimization process of $\boldsymbol{\theta}$ and b, and opens a door to promising avenues.

5.4.4 Domain shift

We introduce a more realistic, cross-domain setting (COCO-20^i to PASCAL-VOC). We argue that such setting is a step towards a more realistic evaluation of these methods, as it can assess the impact on performances caused by a domain shift between the data training distribution and the testing one. We believe that this scenario can be easily found in practice, as even slight alterations in the data collection process might result in a distributional shift. We reproduce the scenario where a large labeled dataset is available (e.g., COCO-20^i), but the evaluation is performed on a target dataset with a different feature distribution (e.g., PASCAL-VOC). As per the original work (Lin *et al.*, 2014), significant differences exist between the two original datasets. For instance, images in MS-COCO have on average 7.7 instances of objects coming from 3.5 distinct categories, while PASCAL-VOC only has an average of 3 instances from 2 distinct categories.

Evaluation. We reuse models trained on each fold of COCO-20^i and generate tasks using images from all the classes in PASCAL-VOC that were not used during training. For instance, fold-0 of this setting means the model was trained on fold-0 of COCO-20^i and tested on the whole Pascal-VOC dataset, after removing the classes seen in training. A complete summary of all the folds is available in the Supplemental material.

Results. We reproduced and compared to the two best performing methods (Tian *et al.*, 2020b; Liu *et al.*, 2020f) using their respective official GitHub repositories. Table 5.4 summarizes the results for the 1-shot and 5-shot cross-domain experiments. We observe that in the presence of domain-shift, our method outperforms existing methods in both 1-shot and 5-shot scenarios.

Table 5.4 Aggregated domain-shift results, averaged over 4 folds, on COCO-20^i to PASCAL-5^i. Best results in bold. Per-fold results are available in the supplementary material

Method	Backbone	1-shot	5-shot
RPMM (Yang *et al.*, 2020a)	ResNet50	49.6	53.8
PFENet (Tian *et al.*, 2020b)		61.1	63.4
RePRI (ours)		**63.1**	**66.2**
Oracle-RePRI	Resnet-50	76.2	79.7

5.4.5 Ablation studies

Impact of each term in the main objective. While Fig. 5.1 provides *qualitative* insights on how each term in Eq. (5.1) affects the final prediction, this section provides a *quantitative* evaluation of their impact, evaluated on PASCAL-5^i (Table 5.5). Quantitative results confirm the qualitative insights observed in Fig. 5.1, as both CE and CE + $\mathcal{H}$ losses drastically degrade the

Table 5.5 Ablation study on the effect of each term in our loss in Eq. (5.1), evaluated on PASCAL-5^i

	1-shot					5-shot				
Loss	Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	Mean
CE	39.7	49.3	37.3	27.5	38.5	56.5	66.4	60.1	49.0	58.0
$\mathrm{CE} + \mathcal{H}$	45.7	61.7	48.2	36.4	48.0	56.8	68.5	61.3	47.0	58.4
$\mathrm{CE} + \mathcal{H} + \mathcal{D}_{KL}$	**59.8**	**68.3**	**62.1**	**48.5**	**59.7**	**64.6**	**71.4**	**71.1**	**59.3**	**66.6**

performance compared to the proportion-regularized loss, i.e., $\mathrm{CE} + \mathcal{H} + \mathcal{D}_{KL}$. For example, in the 1-shot scenario, simply minimizing the CE results in more than 20% of difference compared to the proposed model. In this case, the prototype $\boldsymbol{\theta}$ tends to overfit the support sample and only activates regions of the query object that strongly correlate with the support information. Such behavior hampers the performance when the support and query objects exhibit slights changes in shape or histogram colors, for example, which may be very common in practice. Adding the entropy term $\mathcal{H}$ to CE partially alleviates this problem, as it tends to reinforce the model in being confident on positive pixels initially classified with mid or low confidence. Nevertheless, despite improving the naive CE based model, the gap with the proposed model remains considerably large, with 10% difference. One may notice that the differences between CE and $\mathrm{CE} + \mathcal{H} + \mathcal{D}_{KL}$ decrease in the 5-shot setting, since overfitting 5 support samples simultaneously becomes more difficult. The results from this ablation experiment reinforce our initial hypothesis that the proposed KL term based on the size parameter $\boldsymbol{\pi}$ acts as a strong regularizer.

Influence of parameter $\boldsymbol{\pi}$ misestimation: Precisely knowing the foreground/background (F/B) proportion of the query object is unrealistic. To quantify the deviation from the exact F/B proportion $\boldsymbol{\pi}^*$, we introduce the relative error on the foreground size:

$$\delta^{(t)} = \frac{\boldsymbol{\pi}_1^{(t)}}{\boldsymbol{\pi}_1^*} - 1, \tag{5.5}$$

where $\boldsymbol{\pi}_1^*$ represents the exact foreground proportion in the query image, extracted from its corresponding ground truth, and $\boldsymbol{\pi}_1^{(t)}$ our estimate at iteration t, which is derived from the soft predicted segmentation. As observed from Fig. 5.1, the initial prototype often results in a blurred probability map, from which only a very coarse estimate of the query proportion can be inferred and used as $\boldsymbol{\pi}^{(0)}$. The distribution of δ over 5000 tasks is presented in Fig. 5.2. It clearly shows that the initial prediction typically provides an overestimate of the actual query foreground size, while finetuning the classifier $\boldsymbol{\theta}$ for 10 iterations with our main loss (Eq. 5.1) already provides a strictly more accurate estimate, as conveyed by the right box plot in Fig. 5.2, with an average δ around 0.7. Now, a natural question remains: **how good does the estimate need to be in order to approach the oracle results?** To answer this, we carry out a series of controlled experiments where, instead of computing $\boldsymbol{\pi}^{(t)}$ with Eq. 5.3, we use a δ-perturbed oracle at initialization, such that $\boldsymbol{\pi}_1^{(t)} = \boldsymbol{\pi}_1^*(1+\delta)$. Each point in Fig. 5.2 represents the mIoU obtained over 5000 tasks for a given perturbation δ. Fig. 5.2 reveals that exact F/B proportion is not required to significantly close the gap with the oracle. Specifically, foreground size estimates ranging from -10% to +30% with respect to the oracle proportion are sufficient to achieve 70%+ of mIOU, which represents an improvement of 10% over the current state-of-the art. This suggests that more refined size estimation methods may significantly increase the performance of the proposed method.

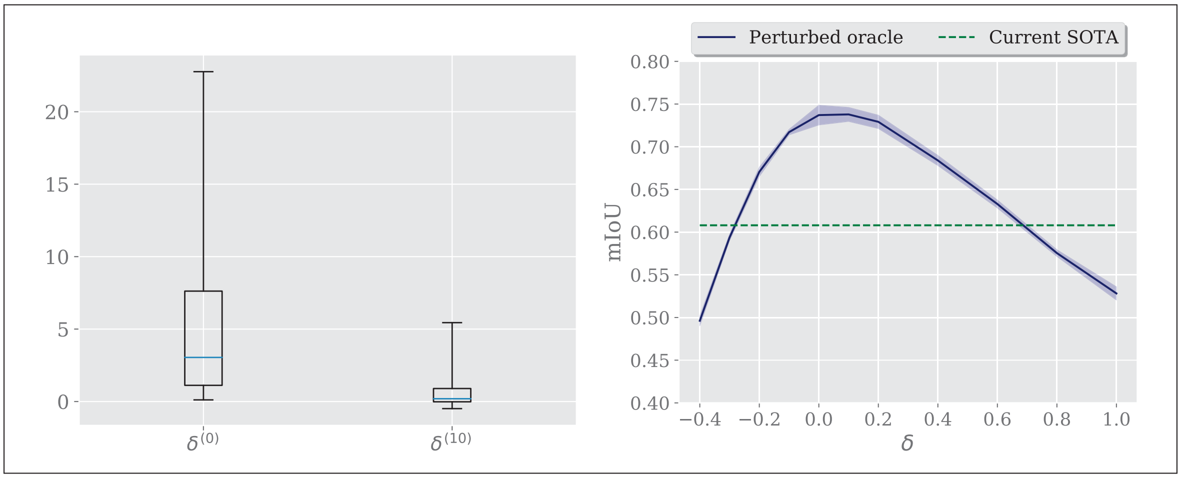

Figure 5.2 Experiments on $\boldsymbol{\pi}$ misestimation. Both figures are computed using 5 runs of 1000 1-shot tasks, each on the fold-0 of PASCAL5^i. (Left): Relative error δ distribution of our current method, at initialization $\delta^{(0)}$ and after 10 gradient iterations $\delta^{(10)}$. (Right) Mean-IoU versus enforced relative foreground size error δ in the parameter $\boldsymbol{\pi}^{(0)}$

Table 5.6 Number of tasks performed per second, and the corresponding mIoU performances on PASCAL-5^i

	1-shot		5-shot	
Method	FPS	mIoU	FPS	mIoU
RPMMS (Yang *et al.*, 2020a)	18.2	51.5	9.4	57.3
PFENet (Tian *et al.*, 2020b)	15.9	60.8	5.1	61.9
RePRI (ours)	12.8	59.7	4.4	66.6

Computational efficiency: We now inspect the computational cost of the proposed solution, and compare it to recent existing methods. Unlike prior work, we solve an optimization problem at inference, which naturally slows down the inference process. However, in our case, only a single prototype vector $\boldsymbol{\theta} \in \mathbb{R}^C$, where we recall C is the feature channel dimension, and a bias $b \in \mathbb{R}$ need to be optimized for each task. Furthermore, in our setting $C = 512$, and therefore the problem can still be solved relatively efficiently, leading to reasonable inference times. In Table 5.6, we summarize the FPS rate at inference for our method, as well as for two competing approaches that only require a forward pass. We can observe that, unsurprisingly, our method reports lower FPS rates, without becoming unacceptably slower. The reported values indicate that the differences in inference times are small compared to, for example, the approach in (Tian *et al.*, 2020b). Particularly, in the 1-shot scenario, our method processes tasks 3 FPS slower than (Tian *et al.*, 2020b), whereas this gap narrows down to 0.7 FPS in the 5-shot setting.

Conclusion

Without resorting to the popular meta-learning paradigm, our proposed RePRI achieves new state-of-the-art results on standard 5-shot segmentation benchmarks, while being on par with best performing approaches in the 1-shot setting. RePRI is modular and can, therefore, be used in conjunction with any feature extractor regardless how the base training was performed. Supported by the findings in this work, we believe that the relevance of the episodic training should be re-considered in the context of few-shot segmentation, and we provide a strong baseline to stimulate future research in this topic. Our results indicate that current state-of-the-art

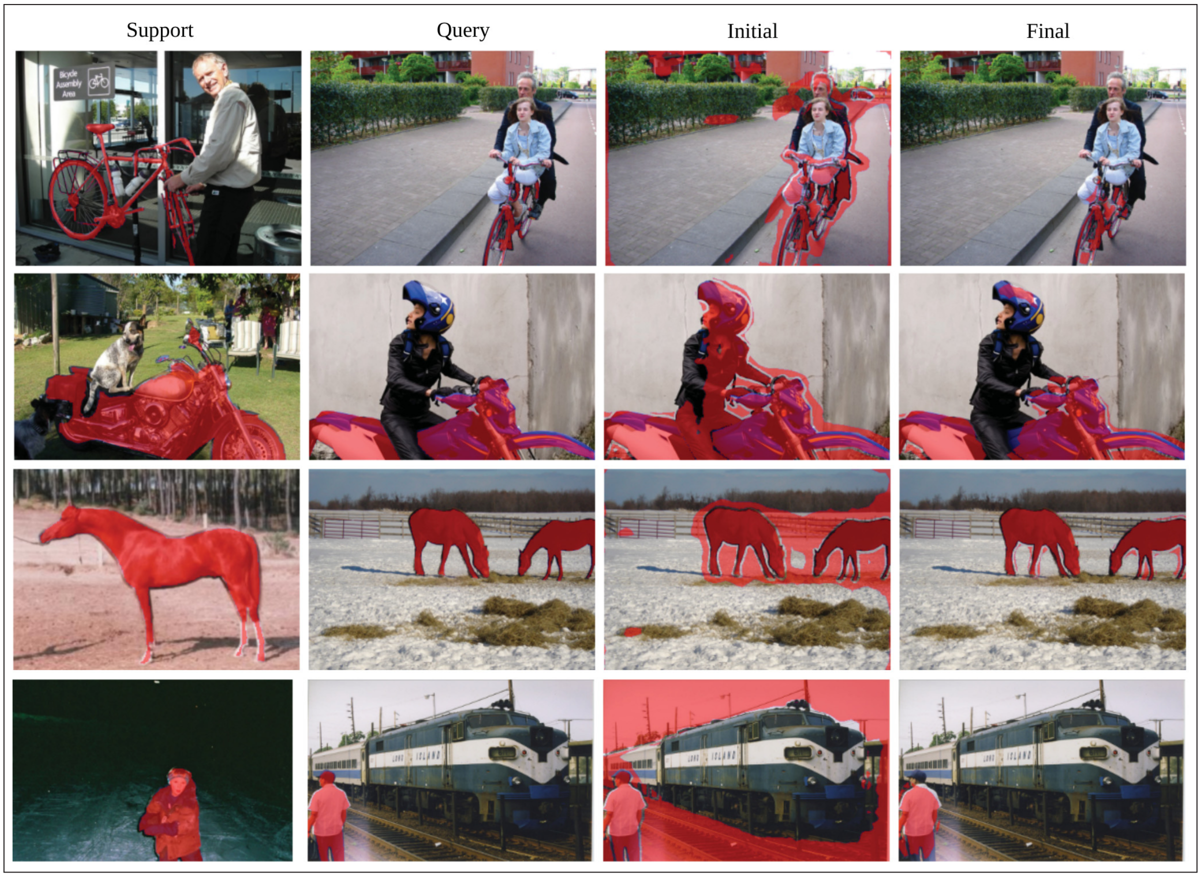

Figure 5.3 Qualitative results. *Initial* column refers to the predictions right after initializing the prototypes, while *Final* column refers to the prediction after running our inference. Best viewed in colors in high resolution

methods may have difficulty with more challenging settings, when dealing with domain shifts or conducting inference on tasks whose structures are different from those seen in training – scenarios that have been completely overlooked in the literature. These findings align with recent observations in few shot classification (Cao *et al.*, 2020; Chen *et al.*, 2019). Furthermore, embedding more accurate foreground-background proportion estimates appears to be a very promising way of constraining the inference, as demonstrated with the significantly improved results obtained by the oracle.

CONCLUSION AND RECOMMENDATIONS

Throughout this thesis, we have developed new methods that aim to efficiently reuse pre-trained models for lowly labeled concepts. To mitigate the lack of labeled data, we have explored transductive learning as a principled framework to extract ad-hoc knowledge from the unlabeled data of interest, thereby solving local problems rather than general ones. Specifically, the works presented in this thesis adapt ideas borrowed from statistics and information theory to utilize unlabeled data, with simple but efficient modifications tailored to each and every task. As such, it is worth noting that while this thesis' main field of application remains computer vision, we anticipate that the findings and recommendations would be applicable to other domains such as natural language and audio processing. We summarize our findings and recommendations in the following three key points.

As a first takeaway, we have empirically demonstrated the potential benefits of resorting to transduction in both classification and semantic segmentation tasks. Specifically, we have shown the power of higher-order terms, such as those derived from marginal distributions' estimates, in regularizing transductive losses and leading to better optima. These impressive results, including those of the proportion oracle on segmentation tasks, provide strong evidence that transduction, along with side meta-data, prior knowledge, or weak labels, constitutes a promising avenue for few-shot learning.

The second role of this thesis was to expose and highlight the challenges that pertain to transduction. In particular, without any prior knowledge about the unlabeled data of interest, transduction can be risky. Distribution properties such as high-class imbalance or the presence of unknown classes can heavily mitigate the potential benefits of transduction, or even worse, lead to much worse performances than simple inductive baselines. Therefore, we advocate for a careful analysis and understanding of the objectives used, with particular attention to non-trivial biases, such as uniform prior, and an explicit consideration and modeling of potential sources of noise, such as the presence of outliers.

Finally, we hope that this thesis can further motivate the development of a model-agnostic few-shot learning paradigm. The foundation models paradigm is gaining ground across several subtasks within vision, including domain adaptation, with the growing popularity of source-free domain adaptation and test-time adaptation settings that explicitly disentangle training from the adaptation procedure. The once dominating meta-learning paradigm, in which training and inference are generally entangled, is gradually being phased out, with the idea that few-shot should be treated as an easily pluggable capacity rather than an intrinsic property of a model. This idea extends beyond computer vision, as evidenced by the recent dramatic adoption of conversational Large Language Models (LLMs) in Natural Language Processing, such as ChatGPT, which positions low-resource black-box adaptation of pretrained models at the heart of current research in natural language processing. The findings of this thesis could benefit emerging families of methods in NLP, such as in-context learning or prompt-tuning. Inference-centered methods that abstract away architectures and pretraining procedures represent a significant step toward the real-world applicability of few-shot learning.

APPENDIX I

SUPPLEMENTARY MATERIAL FOR THE PAPER TITLED *TRANSDUCTIVE INFORMATION MAXIMIZATION FOR FEW-SHOT LEARNING*

1. Mathematical proofs

Proof of Proposition 2.2.0.1

Proof. Let us start from the initial optimization problem:

$$\min_{\boldsymbol{\theta}} \quad \sum_{k=1}^{K} \widehat{p_k} \log \widehat{p_k} - \frac{\alpha}{|\mathbb{Q}|} \sum_{i \in \mathbb{Q}} \sum_{k=1}^{K} p_{ik} \log p_{ik} - \frac{\lambda}{|\mathbb{S}|} \sum_{i \in \mathbb{S}} \sum_{k=1}^{K} y_{ik} \log p_{ik} \tag{A I-1}$$

We can reformulate problem (A I-1) using the ADM approach, i.e., by introducing auxiliary variables $\boldsymbol{q} = [q_{ik}] \in \mathbb{R}^{|\mathbb{Q}| \times K}$ and enforcing equality constraint $\boldsymbol{q} = \boldsymbol{p}$, with $\boldsymbol{p} = [p_{ik}] \in \mathbb{R}^{|\mathbb{Q}| \times K}$, in addition to pointwise simplex constraints:

$$\begin{aligned}
\min_{\boldsymbol{\theta}, \boldsymbol{q}} \quad & \sum_{k=1}^{K} \widehat{q_k} \log \widehat{q_k} - \frac{\alpha}{|\mathbb{Q}|} \sum_{i \in \mathbb{Q}} \sum_{k=1}^{K} q_{ik} \log p_{ik} - \frac{\lambda}{|\mathbb{S}|} \sum_{i \in \mathbb{S}} \sum_{k=1}^{K} y_{ik} \log p_{ik} \\
\text{s.t.} \quad & q_{ik} = p_{ik}, \quad i \in \mathbb{Q}, \quad k \in \{1, \ldots, K\} \\
& \sum_{k=1}^{K} q_{ik} = 1, \quad i \in \mathbb{Q} \\
& q_{ik} \geq 0, \quad i \in \mathbb{Q}, \quad k \in \{1, \ldots, K\}
\end{aligned} \tag{A I-2}$$

We can solve constrained problem (A I-2) with a penalty-based approach, which encourages auxiliary pointwise predictions $\boldsymbol{q}_i = [q_{i1}, \ldots, q_{iK}]$ to be close to our model's posteriors $\boldsymbol{p}_i = [p_{i1}, \ldots, p_{iK}]$. To add a penalty encouraging equality constraints $\boldsymbol{q}_i = \boldsymbol{p}_i$, we use the Kullback–Leibler (KL) divergence, which is given by:

$$\mathcal{D}_{\text{KL}}(\boldsymbol{q}_i || \boldsymbol{p}_i) = \sum_{k=1}^{K} q_{ik} \log \frac{q_{ik}}{p_{ik}}$$

Thus, our constrained optimization problem becomes:

$$\begin{aligned}
\min_{\boldsymbol{\theta},\boldsymbol{q}} \quad & \sum_{k=1}^{K} \widehat{q_k} \log \widehat{q_k} - \frac{\alpha}{|\mathbb{Q}|} \sum_{i \in \mathbb{Q}} \sum_{k=1}^{K} q_{ik} \log p_{ik} - \frac{\lambda}{|\mathbb{S}|} \sum_{i \in \mathbb{S}} \sum_{k=1}^{K} y_{ik} \log p_{ik} + \frac{1}{|\mathbb{Q}|} \sum_{i \in \mathbb{Q}} \mathcal{D}_{\text{KL}}(\boldsymbol{q}_i || \boldsymbol{p}_i) \\
\text{s.t.} \quad & \sum_{k=1}^{K} q_{ik} = 1, \quad i \in \mathbb{Q} \\
& q_{ik} \geq 0, \quad i \in \mathbb{Q}, \quad k \in \{1, \dots, K\}
\end{aligned} \tag{A I-3}$$

□

Proof of Proposition 2.2.0.2

Proof. Recall that we consider a softmax classifier over distances to weights $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K\}$. To simplify the notations, we will omit the dependence upon $\boldsymbol{\phi}$ in what follows, and write $z_i = \frac{f_{\boldsymbol{\phi}}(\boldsymbol{x}_i)}{\|f_{\boldsymbol{\phi}}(\boldsymbol{x}_i)\|}$, such that:

$$p_{ik} = \frac{e^{-\frac{\tau}{2}\|z_i - \boldsymbol{\theta}_k\|^2}}{\sum_{j=1}^{K} e^{-\frac{\tau}{2}\|z_i - \boldsymbol{\theta}_j\|^2}} \tag{A I-4}$$

Without loss of generality, we use $\tau = 1$ in what follows. Plugging the expression of p_{ik} into Eq. (2.4), and grouping terms together, we get:

$$
\begin{aligned}
(2.4) = & \sum_{k=1}^{K} \widehat{q_k} \log \widehat{q_k} - \frac{1+\alpha}{|\mathbb{Q}|} \sum_{i \in \mathbb{Q}} \sum_{k=1}^{K} q_{ik} \log p_{ik} - \frac{\lambda}{|\mathbb{S}|} \sum_{i \in \mathbb{S}} \sum_{k=1}^{K} y_{ik} \log p_{ik} \\
& + \frac{1}{|\mathbb{Q}|} \sum_{i \in \mathbb{Q}} \sum_{k=1}^{K} q_{ik} \log q_{ik}
\end{aligned}
\qquad \text{(A I-5)}
$$

$$
\begin{aligned}
= & \sum_{k=1}^{K} \widehat{q_k} \log \widehat{q_k} \\
& + \frac{1+\alpha}{2|\mathbb{Q}|} \sum_{i \in \mathbb{Q}} \sum_{k=1}^{K} q_{ik} \left\|\boldsymbol{z}_i - \boldsymbol{\theta}_k\right\|^2 + \frac{1+\alpha}{|\mathbb{Q}|} \sum_{i \in \mathbb{Q}} \log \left(\sum_{j=1}^{K} e^{-\frac{1}{2}\left\|\boldsymbol{z}_i - \boldsymbol{\theta}_j\right\|^2} \right) \\
& + \frac{\lambda}{2|\mathbb{S}|} \sum_{i \in \mathbb{S}} \sum_{k=1}^{K} y_{ik} \left\|\boldsymbol{z}_i - \boldsymbol{\theta}_k\right\|^2 + \frac{\lambda}{|\mathbb{S}|} \sum_{i \in \mathbb{S}} \log \left(\sum_{j=1}^{K} e^{-\frac{1}{2}\left\|\boldsymbol{z}_i - \boldsymbol{\theta}_j\right\|^2} \right) \\
& + \frac{1}{|\mathbb{Q}|} \sum_{i \in \mathbb{Q}} \sum_{k=1}^{K} q_{ik} \log q_{ik}
\end{aligned}
\qquad \text{(A I-6)}
$$

Now, we can solve our problem approximately by alternating two sub-steps: one sub-step optimizes w.r.t classifier weights $\boldsymbol{\theta}$ while auxiliary variables $\boldsymbol{q}$ are fixed; another sub-step fixes $\boldsymbol{\theta}$ and update $\boldsymbol{q}$.

- $\boldsymbol{\theta}$-update: Omitting the terms that do not involve $\boldsymbol{\theta}$, Eq. (A I-5) reads:

$$
\begin{aligned}
& \underbrace{\frac{\lambda}{2|\mathbb{S}|} \sum_{i \in \mathbb{S}} y_{ik} \left\|\boldsymbol{z}_i - \boldsymbol{\theta}_k\right\|^2 + \frac{1+\alpha}{2|\mathbb{Q}|} \sum_{i \in \mathbb{Q}} q_{ik} \left\|\boldsymbol{z}_i - \boldsymbol{\theta}_k\right\|^2}_{C\text{:convex}} \\
+ & \underbrace{\frac{\lambda}{|\mathbb{S}|} \sum_{i \in \mathbb{S}} \log \left(\sum_{j=1}^{K} e^{-\frac{1}{2}\left\|\boldsymbol{z}_i - \boldsymbol{\theta}_j\right\|^2} \right) + \frac{1+\alpha}{|\mathbb{Q}|} \sum_{i \in \mathbb{Q}} \log \left(\sum_{j=1}^{K} e^{-\frac{1}{2}\left\|\boldsymbol{z}_i - \boldsymbol{\theta}_j\right\|^2} \right)}_{\bar{C}\text{:non-convex}}
\end{aligned}
\qquad \text{(A I-7)}
$$

One can notice that objective (A I-5) is not convex w.r.t $\boldsymbol{\theta}_k$. Actually, it can be split into convex and non-convex parts as in Eq. (A I-7). Thus, we cannot simply set the gradients to 0 to get the optimal $\boldsymbol{\theta}_k$. The non-convex part can be linearized at current solution $\boldsymbol{\theta}_k^{(t)}$ as follows:

$$\begin{aligned}\bar{C}(\boldsymbol{\theta}_k) &\approx \bar{C}(\boldsymbol{\theta}_k^{(t)}) + \frac{\partial \bar{C}}{\partial \boldsymbol{\theta}_k}(\boldsymbol{\theta}_k^{(t)})^T(\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^{(t)}) \\ &\overset{\mathrm{c}}{=} \frac{\lambda}{|\mathbb{S}|}\sum_{i\in\mathbb{S}} p_{ik}^{(t)}(\boldsymbol{z}_i - \boldsymbol{\theta}_k^{(t)})^T\boldsymbol{\theta}_k + \frac{1+\alpha}{|\mathbb{Q}|}\sum_{i\in\mathbb{Q}} p_{ik}^{(t)}(\boldsymbol{z}_i - \boldsymbol{\theta}_k^{(t)})^T\boldsymbol{\theta}_k \end{aligned} \tag{A I-8}$$

Where $\overset{\mathrm{c}}{=}$ stands for "equal, up to an additive constant". By adding this linear term to the convex part C, we can obtain a strictly convex objective in $\boldsymbol{\theta}_k$, whose gradients w.r.t $\boldsymbol{\theta}_k$ read:

$$\begin{aligned}\frac{\partial(AI-7)}{\partial \boldsymbol{\theta}_k} \approx & \frac{\lambda}{|\mathbb{S}|}[\sum_{i\in\mathbb{S}} y_{ik}(\boldsymbol{z}_i - \boldsymbol{\theta}_k) + p_{ik}^{(t)}(\boldsymbol{z}_i - \boldsymbol{\theta}_k^{(t)})] + \\ & \frac{1+\alpha}{|\mathbb{Q}|}[\sum_{i\in\mathbb{Q}} q_{ik}(\boldsymbol{z}_i - \boldsymbol{\theta}_k) + p_{ik}^{(t)}(\boldsymbol{z}_i - \boldsymbol{\theta}_k^{(t)})]\end{aligned} \tag{A I-9}$$

Note that the approximation we do here is similar in spirit to concave-convex procedures, which are well known in optimization. Concave-convex techniques proceed as follows: for a function in the form of a sum of a concave term and a convex term, the concave part is replaced by its first-order approximation, while the convex part is kept as is. The difference here is that the part that we linearize in Eq. (A I-7) is not concave. Setting the gradients above to 0 yields the optimal solution for the approximate objective.

Another solution to obtain a strictly convex objective would have been to discard the non-convex part $\bar{C}$. Very interestingly, in this case, one would recover $\boldsymbol{\theta}_k$ updates that would very much resemble the prototype updates of the K-means clustering algorithm (slightly modified to take into account the fact that for support points in $\mathbb{S}$ have labels). Note that the link between regularized K-means and mutual information maximization has been extensively explored in (Jabi *et al.*, 2019). Of course, in this case, the approximation is not as good as the first-order approximation above, and we found that omitting the non-convex part might decrease the performances significantly.

- $\boldsymbol{q}$-update: With weights $\boldsymbol{\theta}$ fixed, the objective is convex w.r.t auxiliary variables $\boldsymbol{q}_i$ (sum of linear and convex functions) and the simplex constraints are affine. Therefore, one can minimize this constrained convex problem for each $\boldsymbol{q}_i$ by solving the Karush-Kuhn-Tucker

(KKT) conditions[14]. The KKT conditions yield closed-form solutions for both primal variable $\boldsymbol{q}_i$ and the dual variable (Lagrange multiplier) corresponding to simplex constraint $\sum_{j=1}^{K} q_{ij} = 1$. Interestingly, the negative entropy of auxiliary variables, i.e., $\sum_{k=1}^{K} q_{ik} \log q_{ik}$, which appears in the penalty term, handles implicitly non-negativity constraints $\boldsymbol{q}_i \geq 0$. In fact, this negative entropy acts as a barrier function, restricting the domain of each $\boldsymbol{q}_i$ to non-negative values, which avoids extra dual variables and Lagrangian-dual inner iterations for constraints $\boldsymbol{q}_i \geq 0$. As we will see, the closed-form solutions of the KKT conditions satisfy these non-negativity constraints, without explicitly imposing them. In addition to non-negativity, for each point i, we need to handle probability simplex constraints $\sum_{k=1}^{K} q_{ik} = 1$. Let $\gamma_i \in \mathbb{R}$ denote the Lagrangian multiplier corresponding to this constraint. The KKT conditions correspond to setting the following gradient of the Lagrangian function to zero, while enforcing the simplex constraints:

$$\frac{\partial (2.4)}{\partial q_{ik}} = -\frac{1+\alpha}{|\mathbb{Q}|} \log p_{ik} + \frac{1}{|\mathbb{Q}|}(\log \widehat{q_k} + 1) + \frac{1}{|\mathbb{Q}|}(\log q_{ik} + 1) + \gamma_i \tag{A I-10}$$

$$= \frac{1}{|\mathbb{Q}|}\left(\log(\frac{q_{ik}\widehat{q_k}}{p_{ik}^{1+\alpha}}) + 2\right) + \gamma_i \tag{A I-11}$$

This yields:

$$q_{ik} = \frac{p_{ik}^{1+\alpha}}{\widehat{q_k}} e^{-(\gamma_i|\mathbb{Q}|+2)} \tag{A I-12}$$

Applying simplex constraint $\sum_{j=1}^{K} q_{ij} = 1$ to (A I-12), Lagrange multiplier γ_i verifies:

$$e^{-(\gamma_i|\mathbb{Q}|+2)} = \frac{1}{\sum\limits_{j=1}^{K} \frac{p_{ij}^{1+\alpha}}{\widehat{q_j}}} \tag{A I-13}$$

[14] Note that strong duality holds since the objective is convex and the simplex constraints are affine. This means that the solutions of the (KKT) conditions minimize the objective.

Hence, plugging (A I-13) in (A I-12) yields:

$$q_{ik} = \frac{\frac{p_{ik}^{1+\alpha}}{\widehat{q_k}}}{\sum_{j=1}^{K} \frac{p_{ij}^{1+\alpha}}{\widehat{q_j}}} \tag{A I-14}$$

Using the definition of $\widehat{q_k}$, we can decouple this equation:

$$\widehat{q_k} = \frac{1}{|\mathbb{Q}|} \sum_{i \in \mathbb{Q}} q_{ik} \propto \sum_{i \in \mathbb{Q}} \frac{p_{ik}^{1+\alpha}}{\widehat{q_k}} \tag{A I-15}$$

which implies:

$$\widehat{q_k} \propto \left(\sum_{i \in \mathbb{Q}} p_{ik}^{1+\alpha} \right)^{1/2} \tag{A I-16}$$

Plugging this back in Eq. (A I-14), we get:

$$q_{ik} \propto \frac{p_{ik}^{1+\alpha}}{\left(\sum_{i \in \mathbb{Q}} p_{ik}^{1+\alpha} \right)^{1/2}} \tag{A I-17}$$

Notice that $q_{ik} \geq 0$, hence the solution fulfils the positivity constraint of the original problem.

□

2. TIM algorithms

In this section, we provide the pseudo-code for TIM's inference stage (both TIM-GD and TIM-ADM).

3. Summary figure

We hereby provide a summarizing figure of the training and inference stages used in TIM.

Algorithm-A I-1 TIM-ADM

```
Input  : Pre-trained encoder $f_{\boldsymbol{\phi}}$, Task $\{\mathbb{S}, \mathbb{Q}\}$, # iterations $iter$, Temperature $\tau$, Weights $\{\lambda, \alpha\}$
$z_i \leftarrow \frac{f_{\boldsymbol{\phi}}(\boldsymbol{x}_i)}{\|f_{\boldsymbol{\phi}}(\boldsymbol{x}_i)\|_2}$ , $i \in \mathbb{S} \cup \mathbb{Q}$
$\boldsymbol{\theta}_k \leftarrow \frac{\sum_{i \in \mathbb{S}} y_{ik} z_i}{\sum_{i \in \mathbb{S}} y_{ik}}$ , $k \in \{1, \ldots, K\}$
for $i \leftarrow 0$ to $iter$ do
    $p_{ik} \leftarrow \exp\left(-\frac{\tau}{2} \|\boldsymbol{\theta}_k - z_i\|^2\right)$ , $i \in \mathbb{S} \cup \mathbb{Q}$
    $p_{ik} \leftarrow \frac{p_{ik}}{\sum_{l=1}^{K} p_{il}}$
    $q_{ik} \leftarrow \frac{p_{ik}^{1+\alpha}}{\left(\sum_{i \in \mathbb{Q}} p_{ik}^{1+\alpha}\right)^{1/2}}$ , $i \in \mathbb{Q}$
    $q_{ik} \leftarrow \frac{q_{ik}}{\sum_{l=1}^{K} q_{il}}$
    $\boldsymbol{\theta}_k \leftarrow \frac{\frac{\lambda}{1+\alpha} \sum_{i \in \mathbb{S}} (y_{ik} z_i + p_{ik}(\boldsymbol{\theta}_k - z_i)) + \frac{|\mathbb{S}|}{|\mathbb{Q}|} \sum_{i \in \mathbb{Q}} (q_{ik} z_i + p_{ik}(\boldsymbol{\theta}_k - z_i))}{\frac{\lambda}{1+\alpha} \sum_{i \in \mathbb{S}} y_{ik} + \frac{|\mathbb{S}|}{|\mathbb{Q}|} \sum_{i \in \mathbb{Q}} q_{ik}}$
end for
Result: Query predictions $\hat{y}_i = \arg\max_k p_{ik}$ , $i \in \mathbb{Q}$
```

Algorithm-A I-2 TIM-GD

```
Input  : Pre-trained encoder $f_{\boldsymbol{\phi}}$, Task $\{\mathbb{S}, \mathbb{Q}\}$, # iterations $iter$, Temperature $\tau$, Weights $\{\lambda, \alpha\}$, Learning rate $\gamma$
$z_i \leftarrow \frac{f_{\boldsymbol{\phi}}(\boldsymbol{x}_i)}{\|f_{\boldsymbol{\phi}}(\boldsymbol{x}_i)\|_2}$ , $i \in \mathbb{S} \cup \mathbb{Q}$
$\boldsymbol{\theta}_k \leftarrow \frac{\sum_{i \in \mathbb{S}} y_{ik} z_i}{\sum_{i \in \mathbb{S}} y_{ik}}$ , $k \in \{1, \ldots, K\}$
for $i \leftarrow 0$ to $iter$ do
    $p_{ik} \leftarrow \exp\left(-\frac{\tau}{2} \|\boldsymbol{\theta}_k - z_i\|^2\right)$
    $p_{ik} \leftarrow \frac{p_{ik}}{\sum_{l=1}^{K} p_{il}}$
    $\boldsymbol{\theta}_k \leftarrow \boldsymbol{\theta}_k - \gamma \nabla_{\boldsymbol{\theta}_k} \mathcal{L}_{\text{TIM}}$
end for
Result: Query predictions $\hat{y}_i = \arg\max_k p_{ik}$ , $i \in \mathbb{Q}$
```

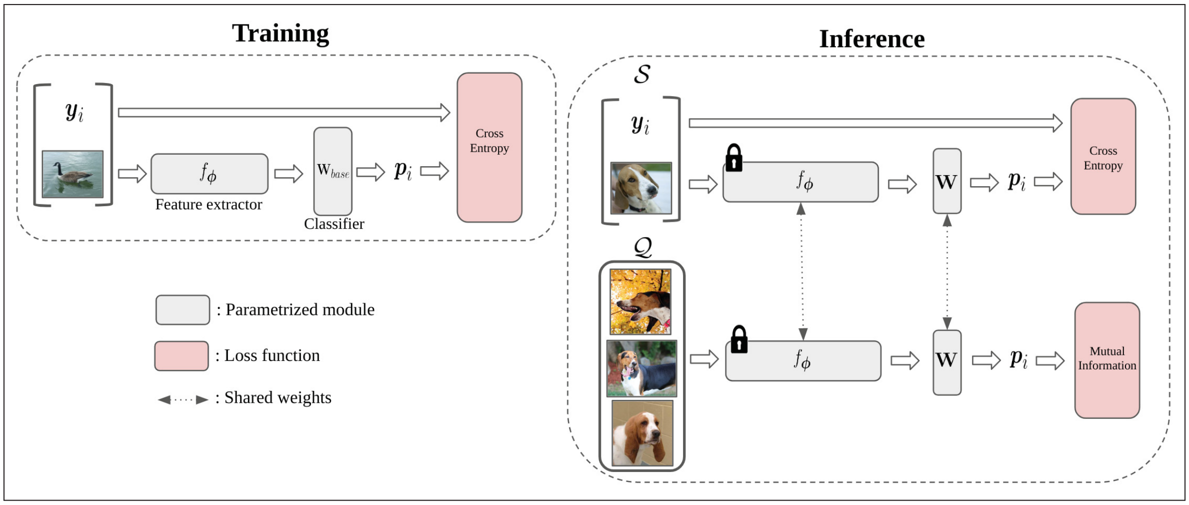

Figure-A I-1 Outline of TIM framework (best viewed in color). First, the feature extractor is trained with the standard cross-entropy on the base classes. Then, it is kept fixed at inference and weights $\boldsymbol{\theta}$ are optimized for by minimizing the cross-entropy on the support set $\mathbb{S}$, while maximizing the mutual information between features and predictions on the query set $\mathbb{Q}$

4. Details of ADM ablation

In Table I-1, we provide the $\boldsymbol{\theta}$ and $\boldsymbol{q}$ updates for each configuration of the TIM-ADM ablation study, whose results were presented in Table 2.4. The proof for each of these updates is very similar to the proof of Proposition 2.2.0.2 detailed in section 1. Therefore, we do not detail it here.

Loss	$\boldsymbol{\theta}_k$ **update**	q_{ik} **update**
CE	$\dfrac{\sum_{i\in\mathbb{S}} y_{ik}\mathbf{z}_i}{\sum_{i\in\mathbb{S}} y_{ik}}$	N/A
$\mathrm{CE} + \widehat{\mathcal{H}}(Y_{\mathbb{Q}}\|X_{\mathbb{Q}})$	-	$\propto p_{ik}^{1+\alpha}$
$\mathrm{CE} - \widehat{\mathcal{H}}(Y_{\mathbb{Q}})$	-	$\propto \dfrac{p_{ik}}{\left(\sum_{i\in\mathbb{Q}} p_{ik}\right)^{1/2}}$
$\mathrm{CE} - \widehat{\mathcal{H}}(Y_{\mathbb{Q}}) + \widehat{\mathcal{H}}(Y_{\mathbb{Q}}\|X_{\mathbb{Q}})$	-	-

Table-A I-1 The **W** and ***q***-updates for each case of the ablation study. "-" refers to the updates in Proposition 2.2.0.2. "NA" refers to non-applicable

APPENDIX II

SUPPLEMENTARY MATERIAL FOR THE PAPER TITLED *REALISTIC EVALUATION OF TRANSDUCTIVE FEW-SHOT LEARNING*

1. On the performance of α-TIM on the standard balanced setting

In the main tables of the paper, we did not include the performances of α-TIM in the standard balanced setting. Here, we emphasize that α-TIM is a generalization of TIM (Boudiaf *et al.*, 2020a) as when $\alpha \to 1$ (i.e., the α-entropies tend to the Shannon entropies), α-TIM tends to TIM. Therefore, in the standard setting, where optimal hyper-parameter α is obtained over validation tasks that are balanced (as in the standard validation tasks of the original TIM and the other existing methods), the performance of α-TIM is the same as TIM. When α is tuned on balanced validation tasks, we obtain an optimal value of α very close to 1, and our α-mutual information approaches the standard mutual information. When the validation tasks are uniformly random, as in our new setting and in the validation plots we provided in the main figure, one can see that the performance of α-TIM remains competitive when we tend to balanced testing tasks (i.e., when a is increasing), but is significantly better than TIM when we tend to uniformly-random testing tasks ($a = 1$). These results illustrate the flexibility of α-divergences, and are in line with the technical analysis provided in the main paper.

2. Comparison with DeepEMD

The recent method (Zhang *et al.*, 2020a) achieves impressive results in the inductive setting. As conveyed in the main paper, inductive methods tend to be unaffected by class imbalance on the query set, which legitimately questions whether strong inductive methods should be preferred over transductive ones, including our proposed α-TIM. In the case of DeepEMD, we expand below on the levers used to obtain such results, and argue those are orthogonal to the loss function, and therefore to our proposed α-TIM method. More specifically:

1. DeepEMD uses richer feature representations: While all the methods we reproduce use the standard global average pooling to obtain a single feature vector per image, DeepEMD-FCN

leverages dense feature maps (i.e without the pooling layer). This results in a richer, much higher-dimensional embeddings. For instance, the standard RN-18 yields a 512-D vector per image, while the FCN RN-12 used by DeepEMD yields a 5x5x640-D feature map (i.e 31x larger). As for DeepEMD-Grid and DeepEMD-Sampling, they build feature maps by concatenating feature extracted from N different patches taken from the original image (which requires as many forward passes through the backbone). Also, note that prototypes optimized for during inference have the same dimension as the feature maps. Therefore, taking richer and larger feature representations also means increasing the number of trainable parameters at inference by the same ratio.

2. DeepEMD uses a more sophisticated notion of distance (namely the Earth Moving Distance), introducing an EMD layer, different from the standard classification layer. While all methods we reproduced are based on simple and easy-to-compute distances between each feature and the prototypes (e.g Euclidian, dot-product, cosine distance), the flow-based distance used by DeepEMD captures more complex patterns than the usual Euclidian distance, but is also much more demanding computationally (as it requires solving an LP program).

Now, we want to emphasize that the model differences mentioned above can be straightforwardly applied to our α-TIM (and likely the other methods) in order to boost the results at the cost of a significant increase of compute requirement. To demonstrate this point, we implemented our α-TIM in with the three ResNet-12 based architectures proposed in DeepEMD (cf Table II-1) using our imbalanced tasks, and consistently observed +3 to +4 without changing any optimization hyper-parameter from their setting, and using the pre-trained models the authors have provided. This figure matches the improvement observed w.r.t to SimpleShot with the standard models (RN-18 and WRN).

3. Relation between α-entropy and α-divergence

We provide the derivation of Eq. (4) in the main paper, which links α-entropy $\mathcal{H}_\alpha(\boldsymbol{p})$ to the α-divergence:

Table-A II-1 Comparison with DeepEMD (Zhang *et al.*, 2020a). Input: **W**=Whole images are used as input ; **P** = Multiples patches of the whole image are used as input. Embeddings: **G**=Global averaged features are used (i.e 1 feature vector per image) ; **L** = Local features are used (i.e 1 feature map per image)

Method	**Distance**	**RN-18 (W/G)**	**WRN (W/G)**	**FCN RN-12 (W/L)**	**Grid RN-12 (P/L)**	**Sampling RN-12 (P/L)**
SimpleShot (Wang *et al.*, 2019b)	Euclidian	63.0	66.2	—	—	—
α-TIM	Euclidian	67.4	69.8	—	—	—
DeepEMD (Zhang *et al.*, 2020a)	EMD	—	—	65.9	67.8	68.8
α-TIM	EMD	—	—	68.9	72.0	72.6

$$\begin{aligned}\log_\alpha(K) - K^{1-\alpha}\mathcal{D}_\alpha(\boldsymbol{p}||\mathbf{u}_K) &= \frac{1}{1-\alpha}\left(K^{1-\alpha}-1\right) - \frac{K^{1-\alpha}}{\alpha-1}\left(\sum_{k=1}^{K} p_k^\alpha \left(\frac{1}{K}\right)^{1-\alpha} - 1\right)\\ &= \frac{1}{1-\alpha}K^{1-\alpha} - \frac{1}{1-\alpha} - \frac{1}{\alpha-1}\sum_{k=1}^{K} p_k^\alpha + \frac{K^{1-\alpha}}{\alpha-1}\\ &= \frac{1}{\alpha-1}\left(1 - \sum_{k=1}^{K} p_k^\alpha\right) \qquad \text{(A II-1)}\end{aligned}$$

4. Comparison with other types of imbalance

The study in (Ochal *et al.*, 2021) examined the effect of class imbalance on the support set after defining several processes to generate class-imbalanced support sets. In particular, the authors proposed *linear* and *step* imbalance. In a 5-way setting, a typical *linearly* imbalanced few-shot support would look like $\{1, 3, 5, 7, 9\}$ (keeping the total number of support samples equivalent to standard 5-ways 5-shot tasks), while a *step* imbalance task could be $\{1, 9, 9, 9\}$. To provide intuition as to how these two types of imbalance related to our proposed Dirichlet-based sampling scheme, we super-impose Dirichlet's density on all valid *linear* and *step* imbalanced distributions for 3-ways tasks in Figure II-1. Combined, *linear* and *step* imbalanced valid distributions allow to cover a fair part of the simplex, but Dirichlet sampling allows to sample more diverse and arbitrary class ratios.

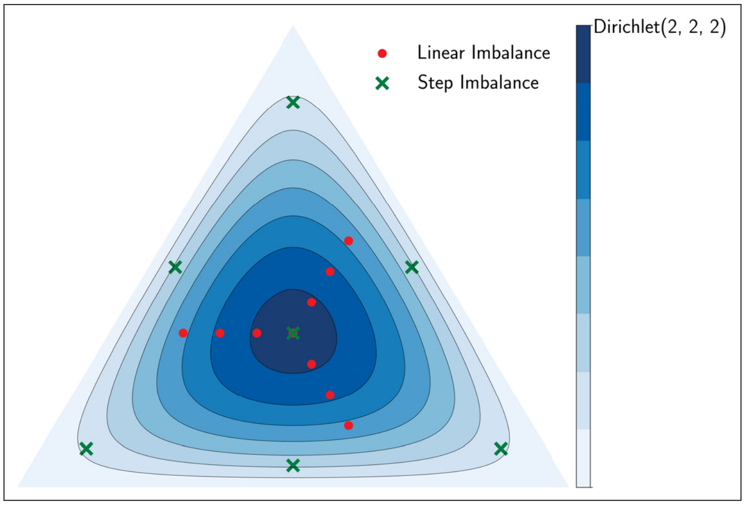

Figure-A II-1 Comparison of Dirichlet sampling with linear and step imbalance

5. Influence of each term in TIM and α-TIM

We report a comprehensive ablation study, evaluating the benefits of using the α-entropy instead of the Shannon entropy (both conditional and marginal terms), as well as the effect of the marginal-entropy terms in the loss functions of TIM and α-TIM. The results are reported in Table II-2. α-TIM yields better performances in all settings.

On the α-conditional entropy: The results of α-TIM obtained by optimizing the conditional entropy alone (without the marginal term) are 4.5 to 7.2% higher in 1-shot, 0.8 to 3.5% higher in 5-shot and 0.1 to 1.3% higher in 10-shot scenarios, in comparison to its Shannon-entropy counterpart in TIM. Note that, for the conditional-probability term, the α-entropy formulation has a stronger effect in lower-shot scenarios (1-shot and 5-shot). We hypothesize that this is due to the shapes of the α-entropy functions and their gradient dynamics (see Fig. 2 in the main paper), which, during training, assigns more weight to confident predictions near the vertices of the simplex ($p = 1$ or $p = 0$) and less weight to uncertain predictions at the middle of the simplex ($p = 0.5$). This discourages propagation of errors during training (i.e., learning from uncertain predictions), which are more likely to happen in lower-shot regimes.

Flexibility of the α-marginal entropy: An important observation is that the marginal-entropy term does even hurt the performances of TIM in the higher shot scenarios (10-shot), even though the results correspond to the best λ over the validation set. We hypothesize that this is due to the strong class-balance bias in the Shannon marginal entropy. Again, due to the shapes of the α-entropy functions and their gradient dynamics, α-TIM tolerates better class imbalance. In the 10-shot scenarios, the performances of TIM decrease by 1.8 to 3.2% when including the marginal entropy, whereas the performance of α-TIM remains approximately the same (with or without the marginal-entropy term). These performances demonstrate the flexibility of α-TIM.

Table-A II-2 An ablation study evaluating the benefits of using the α-entropy instead of the Shannon entropy (both conditional and marginal terms), as well as the effect of the marginal-entropy terms in the loss functions of TIM and α-TIM

Loss	**Dataset**	**Network**	**Method**	**1-shot**	**5-shot**	**10-shot**
$\mathrm{CE} + \mathcal{H}(Y_Q \vert X_Q)$	*mini*-Imagenet	RN-18	TIM	42.2	79.5	85.5
			α-TIM	**48.4**	**82.4**	**86.0**
		WRN	TIM	52.8	82.7	87.5
			α-TIM	**57.3**	**84.6**	**88.0**
	tiered-Imagenet	RN-18	TIM	52.4	83.7	88.4
			α-TIM	**59.0**	**86.3**	**89.2**
		WRN	TIM	49.6	84.1	89.1
			α-TIM	**56.8**	**87.6**	**90.4**
	CUB	RN-18	TIM	56.4	89.0	92.2
			α-TIM	**63.2**	**89.8**	**92.3**
$\mathrm{CE} + \mathcal{H}(Y_Q \vert X_Q) - \mathcal{H}(Y_Q)$	*mini*-Imagenet	RN-18	TIM	67.3	79.8	82.3
			α-TIM	**67.4**	**82.5**	**85.9**
		WRN	TIM	69.8	82.3	84.5
			α-TIM	69.8	**84.8**	**87.9**
	tiered-Imagenet	RN-18	TIM	74.1	84.1	86.0
			α-TIM	**74.4**	**86.6**	**89.3**
		WRN	TIM	75.8	85.4	87.3
			α-TIM	**76.0**	**87.8**	**90.4**
	CUB	RN-18	TIM	74.8	86.9	89.5
			α-TIM	**75.7**	**89.8**	**92.3**

6. Hyper-parameters validation

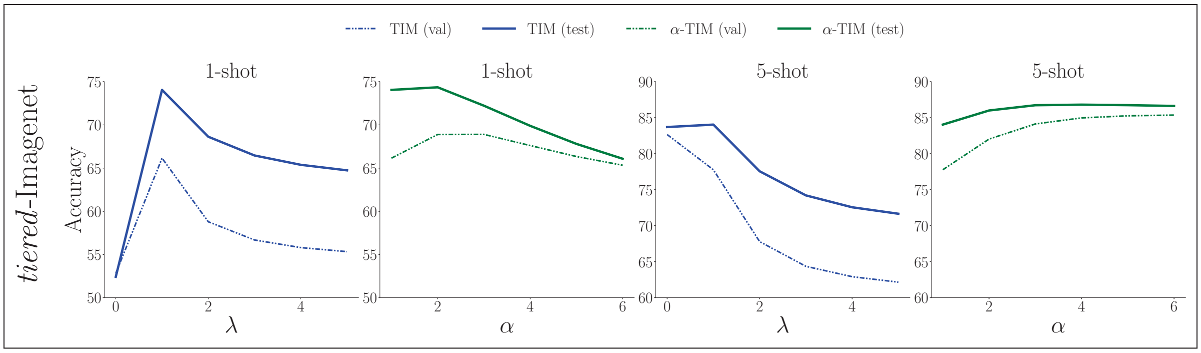

Figure-A II-2 Validation and Test accuracy versus λ for TIM (Boudiaf *et al.*, 2020a) and versus α for α-TIM, using our task-generation protocol. Results are obtained with a RN-18. Best viewed in color

7. Code – Implementation of our framework

As mentioned in our main experimental section, all the methods have been reproduced in our common framework, except for SIB[15] (Hu *et al.*, 2020) and LR-ICI[16] (Wang *et al.*, 2020b), for which we used the official public implementations of the works.

[15] SIB public implementation: https://github.com/hushell/sib_meta_learn

[16] LR-ICI public implementation: https://github.com/Yikai-Wang/ICI-FSL

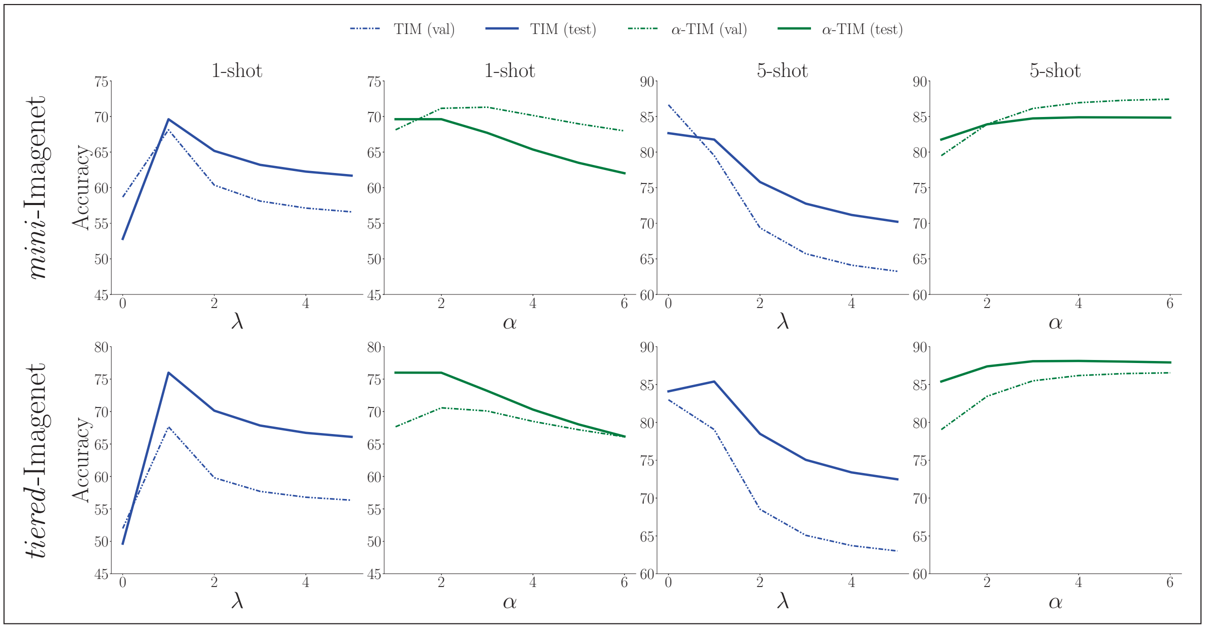

Figure-A II-3 Validation and Test accuracy versus λ for TIM (Boudiaf *et al.*, 2020a) and versus α for α-TIM, using our task-generation protocol. Results are obtained with a WRN. Best viewed in color

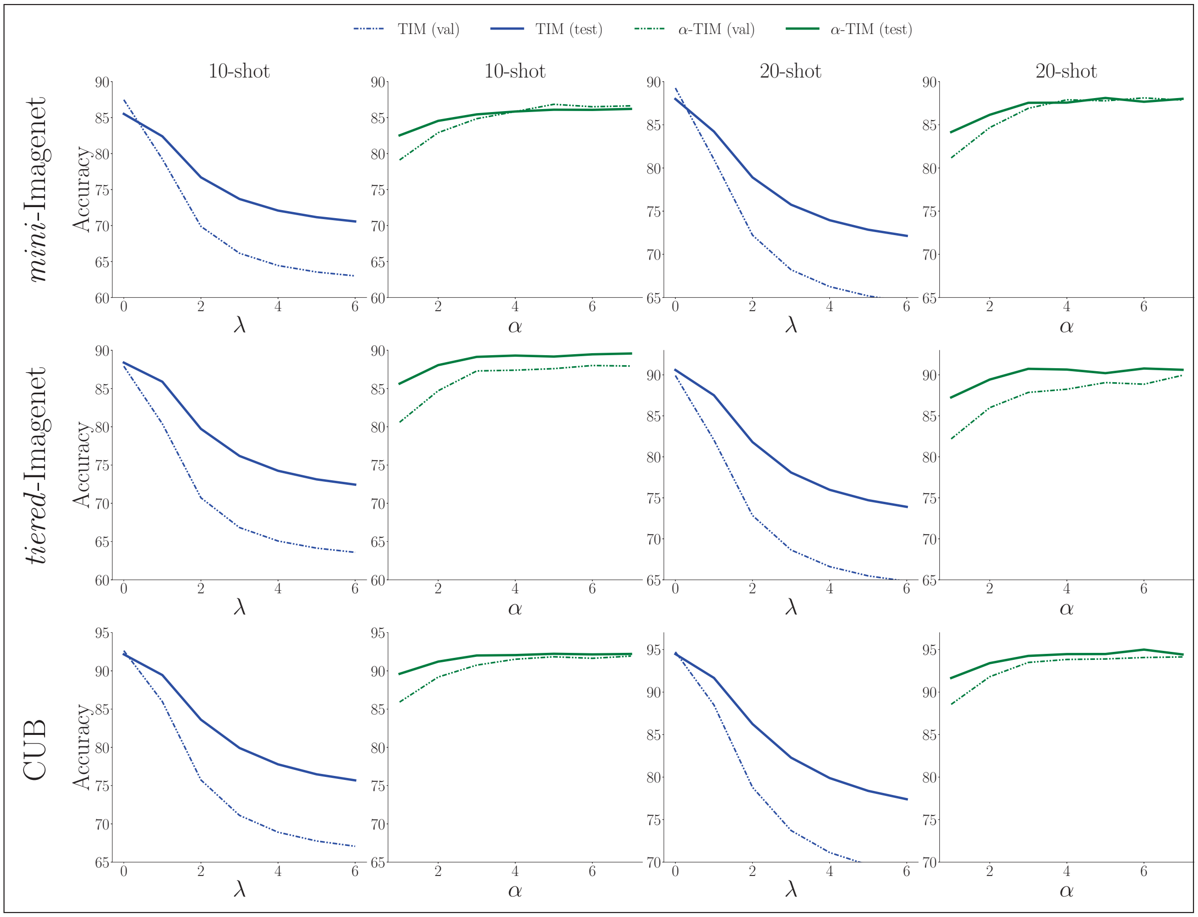

Figure-A II-4 Validation and Test accuracy versus λ for TIM (Boudiaf *et al.*, 2020a) and versus α for α-TIM on 10-shot and 20-shot tasks, using our task-generation protocol. Results are obtained with a RN-18. Best viewed in color

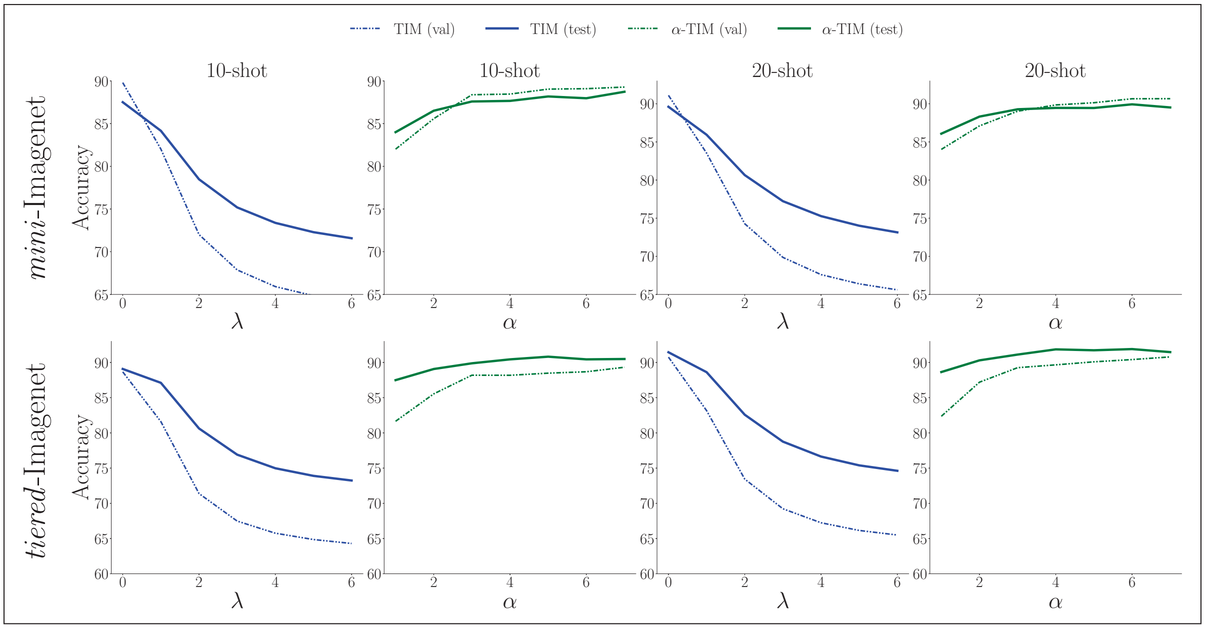

Figure-A II-5 Validation and Test accuracy versus λ for TIM (Boudiaf *et al.*, 2020a) and versus α for α-TIM on 10-shot and 20-shot tasks, using our task-generation protocol. Results are obtained with a WRN. Best viewed in color

APPENDIX III

SUPPLEMENTARY MATERIAL FOR THE PAPER TITLED *MODEL-AGNOSTIC FEW-SHOT OPEN-SET RECOGNITION*

1. Metrics

Here we provide some details about the metrics used in Section 4.5

Acc: the classification accuracy on the closed-set instances of the query set (i.e. $y^q \in \mathbb{C}_{\mathbb{S}}$).

AUROC: the area under the ROC curve is an almost mandatory metric for any OOD detection task. For a set of outlier predictions in $[O, 1]$ and their ground truth (0 for inliers, 1 for outliers), any threshold $\gamma \in [O, 1]$ gives a true positive rate $TP(\gamma)$ (i.e. recall) and a false positive rate $FP(\gamma)$. By rolling this threshold, we obtain a plot of TP as a function of FP i.e. the ROC curve. The area under this curve is a measure of the discrimination ability of the outlier detector. Random predictions lead to an AUROC of 50%.

AUPR: the area under the precision-recall (PR) curve is also a common metric in OOD detection. With the same principle as the ROC curve, the PR curve plots the precision as a function of the recall. Random predictions lead to an AUPR equal to the proportion of outliers in the query set i.e. 50% in our set-up.

Prec@0.9: the precision at 90% recall is the achievable precision on the few-shot open-set recognition task when setting the threshold allowing a recall of 90% for the same task. While AUROC and AUPR are global metrics, *Prec@0.9* measures the ability of the detector to solve a specific problem, which is the detection of almost all outliers (e.g. for raising an alert when open-set instances appear so a human operator can create appropriate new classes). Since all detectors are able to achieve high recall with a sufficiently permissive threshold γ, an excellent way to compare them is to measure the precision of the predictor at a given level of recall (i.e. the proportion of false alarms that the human operator will have to handle). Random predictions lead to a *Prec@0.9* equal to the proportion of outliers in the query set i.e. 50% in our set-up.

Table-A III-1 **Standard Benchmarking**. Evaluating different families of methods on the Few-Shot Open-Set problem, on the popular *tiered*-ImageNet, using a ResNet-12. For each column, a light-gray standard deviation is indicated, corresponding to the maximum deviation observed across methods for that metric. Best methods are shown in bold. Results for PEELER* are reported from (Jeong *et al.*, 2021)

Strategy	Method	1-shot				5-shot			
		Acc	AUROC	AUPR	Prec@0.9	Acc	AUROC	AUPR	Prec@0.9
		±0.74	±0.76	±0.71	±0.52	±0.52	±0.68	±0.75	±0.57
OOD detection	*k*-NN	-	74.62	73.99	61.1	-	80.32	80.15	65.24
	IForest	-	55.03	54.56	51.91	-	62.46	61.32	54.53
	OCVSM	-	71.72	71.98	58.68	-	70.85	67.93	60.88
	PCA	-	68.78	67.74	57.68	-	76.37	76.55	61.5
	COPOD	-	50.99	52.05	51.1	-	52.53	53.32	51.34
	HBOS	-	57.77	57.0	53.1	-	61.06	60.02	54.07
Inductive classifiers	SimpleShot (Wang *et al.*, 2019b)	70.52	70.39	68.42	58.99	84.65	77.67	76.24	63.5
	Baseline ++ (Chen *et al.*, 2019)	70.53	70.34	68.32	59.03	84.78	74.01	72.47	61.25
	FEAT (Ye *et al.*, 2020b)	70.15	52.43	56.44	50.0	83.79	53.31	59.81	50.0
Inductive Open-Set	PEELER* (Liu *et al.*, 2020b)	69.51	65.20	-	-	84.10	73.27	-	-
	SnatcherF (Jeong *et al.*, 2021)	70.15	74.51	73.94	61.01	83.79	81.97	81.65	66.78
	OpenMax (Bendale & Boult, 2016)	70.52	72.71	72.6	59.75	**85.44**	77.94	78.48	62.86
	PROSER (Zhou *et al.*, 2021)	68.96	70.61	70.73	57.99	82.87	75.8	76.66	60.71
Transductive classifiers	LaplacianShot (Ziko *et al.*, 2020)	**76.19**	58.39	58.69	53.96	**85.77**	63.66	63.61	55.11
	BDCSPN (Liu *et al.*, 2020c)	74.80	62.58	62.23	54.92	**85.30**	67.43	67.49	56.25
	TIM-GD (Boudiaf *et al.*, 2020a)	72.89	68.46	66.37	58.24	**85.38**	74.71	73.02	61.69
	PT-MAP (Hu *et al.*, 2021)	71.39	64.86	63.39	56.57	82.66	71.08	69.65	59.14
	LR-ICI (Wang *et al.*, 2020b)	74.18	45.04	48.73	49.85	84.27	45.66	50.02	49.98
Transductive Open-Set	OSLO (ours)	74.32	**79.00**	**79.11**	**64.04**	**85.50**	**87.87**	**88.24**	**73.08**

2. Additional results

Table III-1 shows the benchmark results on *tiered*-ImageNet, and exhibits the same trends observed on *mini*-ImageNet in Section 4.5. Furthermore, we provide a more complete version of Fig. 4.2 in Fig. III-1 and III-2, showing the additional Prec@0.9 metric, along with the results on the WRN2810 provided by (Ye *et al.*, 2020b).

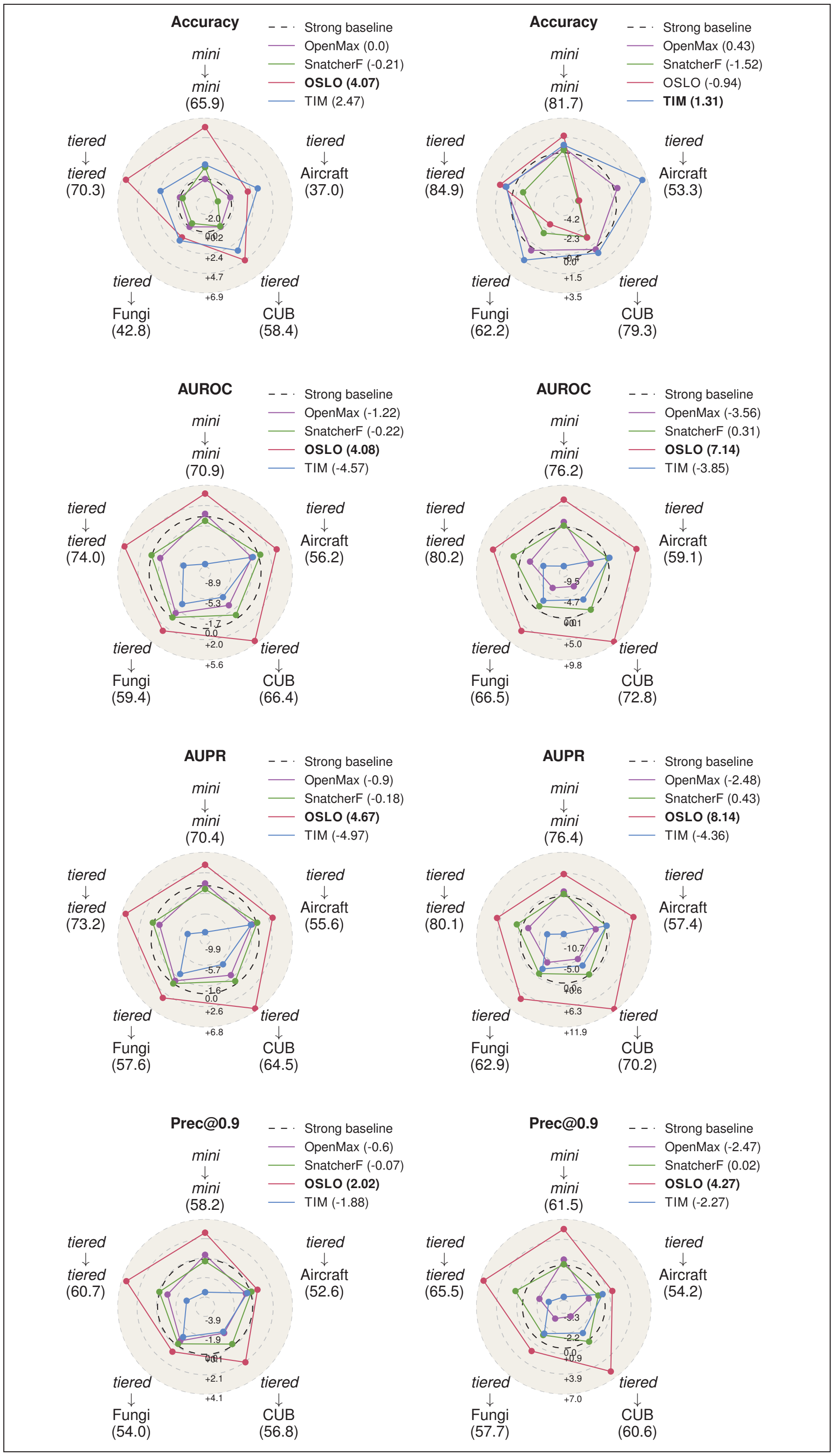

Figure-A III-1 Complete version of Fig. 4.2 with a ResNet-12. (Left column): 1-shot. (Right column): 5-shot

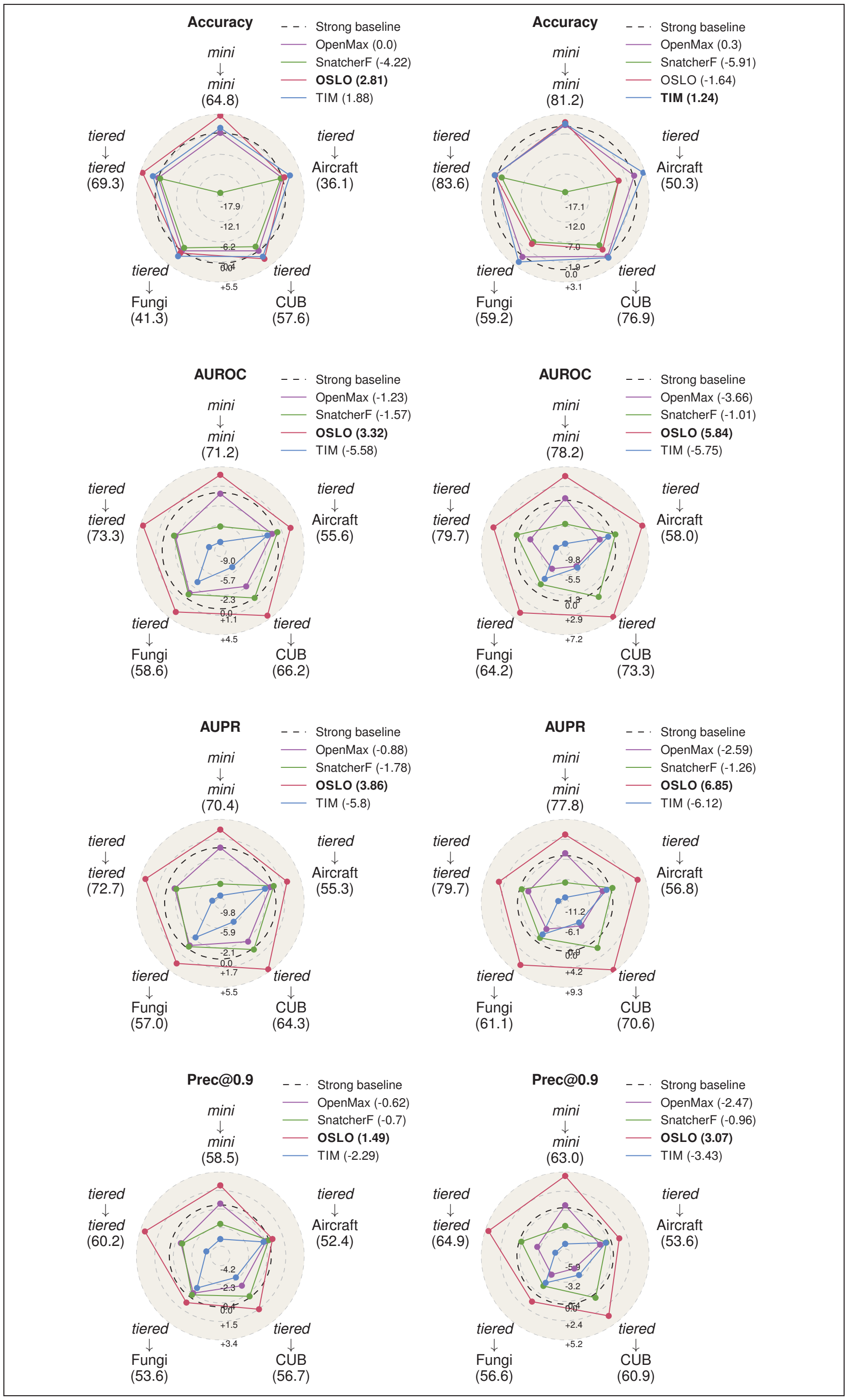

Figure-A III-2 Complete version of Fig. 4.2 with a WideResNet 28-10. (Left column): 1-shot. (Right column): 5-shot. SnatcherF was not included in this plot because a yet misdiagnosed problem occurred with the provided *tiered*-ImageNet checkpoint

APPENDIX IV

SUPPLEMENTARY MATERIAL FOR THE PAPER TITLED *FEW-SHOT SEGMENTATION WITHOUT META-LEARNING: A GOOD TRANSDUCTIVE INFERENCE IS ALL YOU NEED?*

1. Domain shift experiment

In Table IV-1, we show the details of the cross-domain folds used for the domain-shift experiments. Also, in Table IV-2, the per-fold results of the same experiment are available.

Table-A IV-1 Cross-domain folds

	Dataset	Fold 0	Fold 1	Fold 2	Fold 3
Excluded from training	MS-COCO	Person, Airplane, Boat, P. meter, Dog, Elephant Backpack, Suitcase, S. ball, Skateboard, W. glass, Spoon, Sandwich, Hot dog, Chair, D. table, Mouse, Microwave, Fridge, Scissors	Bus, T. light, Bicycle, Bench, Horse, Bear, Umbrella, Frisbee, Kite, Surfboard, Cup, Bowl, Orange, Pizza, Couch, Toilet, Remote, Oven, Book, Teddy	Car, Fire H., Bird, Train, Sheep, Zebra Handbag, Skis, B. bat, T. racket, Fork, Banana, Boroccoli, Donut, P.plant, TV, Keyboard, Toaster, Clock, Hairdrier	Motorcycle, Stop, Cat, Truck , Cow, Giraffe, Tie, Snowboard, B.glove, Bottle, Knife, Apple, Carrot, Cake, Bed, Laptop, Cellphone, Sink, Vase, Toothbrush
Test classes	PASCAL-VOC	Airplane, boat, chair, D. table, Dog, Person	Horse, Sofa, Bicycle, Bus	Bird, Car, P.plant Sheep, Train, TV	Bottle, Cat, Cow, Motorcycle

Table-A IV-2 Per-fold domain-shift results on COCO-20^i to PASCAL-5^i experiment. Best results in bold

		1 shot					5 shot				
Method	Backbone	Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	Mean
RPMM (Yang *et al.*, 2020a) (ECCV'20)	ResNet50	36.3	55.0	52.5	54.6	49.6	40.2	58.0	55.2	61.8	53.8
PFENet (Tian *et al.*, 2020b)		43.2	**65.1**	**66.5**	69.7	61.1	45.1	66.8	68.5	**73.1**	63.4
RePRI (ours)		**52.2**	64.3	64.8	**71.6**	**63.2**	**56.5**	**68.2**	**70.0**	**76.2**	**67.7**
Oracle-RePRI	ResNet50	69.6	71.7	77.6	86.2	76.2	73.5	74.9	82.2	88.1	79.7

2. Results of the 10-shot experiments

In Table IV-3, we give the per-fold results of the 10-shot experiments.

Table-A IV-3 Per-fold 10-shots results on PASCAL-5^i and COCO-20^i. Best results in bold

		PASCAL-5^i					COCO-20^i				
Method	Backbone	Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	Mean
RPMM (Yang *et al.*, 2020a) (ECCV'20)	ResNet50	56.1	68.2	53.9	52.3	57.6	30.9	39.2	28.2	34.0	33.1
PFENet (Tian *et al.*, 2020b)		63.1	70.6	56.6	58.2	62.1	36.9	43.9	38.9	39.1	39.7
RePRI (ours)		**65.7**	**71.9**	**73.3**	**61.2**	**68.1**	**41.6**	**48.2**	**42.1**	**44.5**	**44.1**
Oracle-RePRI	ResNet50	75.6	81.0	82.1	75.6	78.6	57.5	64.7	56.6	56.1	58.7

3. Qualitative results

In Figure IV-1, we provide some qualitative results on PASCAL-5^i that show how our method helps refining the initial predictions of the classifier.

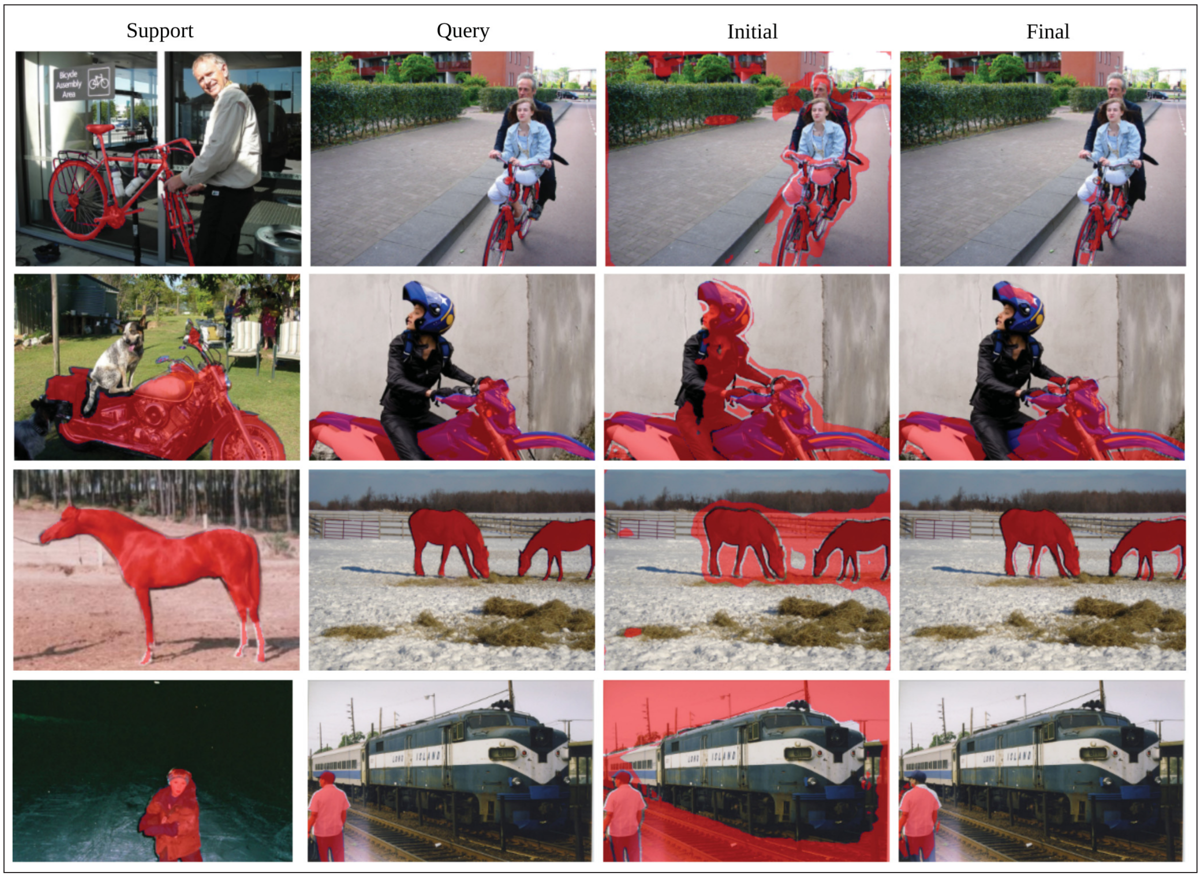

Figure-A IV-1 Qualitative results on PASCAL 5^i. *Initial* column refers to the predictions right after initializing the prototypes, while *Final* column refers to the prediction after running our inference. Best viewed in colors in high resolution

BIBLIOGRAPHY

Amari, S.-I. (2000). α-Divergence Is Unique, Belonging to Both f-Divergence and Bregman Divergence Classes. *IEEE Transactions on Information Theory*, 55(11), 4925–4931.

Antoniou, A., Edwards, H. & Storkey, A. (2019). How to train your MAML.

Arimoto, S. (1977). Information measures and capacity of order α for discrete memoryless channels.

Barlow, H. B. (1989). Unsupervised learning. *Neural Comput.*

Bendale, A. & Boult, T. E. (2016). Towards open set deep networks. *Computer Vision and Pattern Recognition Conference (CVPR).*

Bennequin, E., Bouvier, V., Tami, M., Toubhans, A. & Hudelot, C. (2021). Bridging Few-Shot Learning and Adaptation: New Challenges of Support-Query Shift.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A. & Raffel, C. A. (2019). Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems (NeurIPS).*

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E. et al. (2021). On the opportunities and risks of foundation models.

Boudiaf, M., Masud, Z. I., Rony, J., Dolz, J., Piantanida, P. & Ayed, I. B. (2020a). Transductive Information Maximization For Few-Shot Learning. *Neural Information Processing Systems (NeurIPS).*

Boudiaf, M., Rony, J., Ziko, I. M., Granger, E., Pedersoli, M., Piantanida, P. & Ayed, I. B. (2020b). Metric learning: cross-entropy vs. pairwise losses.

Boudiaf, M., Kervadec, H., Masud, Z. I., Piantanida, P., Ben Ayed, I. & Dolz, J. (2021). Few-Shot segmentation without Meta-Learning: A good transductive inference is all you need? *Computer Vision and Pattern Recognition Conference (CVPR).*

Boyd, S., Parikh, N., Chu, E., Peleato, B. & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*.

Boykov, Y. & Funka-Lea, G. (2006). Graph cuts and efficient ND image segmentation. *International journal of computer vision*, 70(2), 109–131.

Bronskill, J., Gordon, J., Requeima, J., Nowozin, S. & Turner, R. (2020). Tasknorm: Rethinking batch normalization for meta-learning. *International Conference on Machine Learning (ICML)*, pp. 1153–1164.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020). Language models are few-shot learners.

Burns, C. & Steinhardt, J. (2021). Limitations of Post-Hoc Feature Alignment for Robustness. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cao, T., Law, M. & Fidler, S. (2020). A Theoretical Analysis of the Number of Shots in Few-Shot Learning. *International Machine Learning Society (ICML)*.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P. & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. *ICCV*.

Chen, W.-Y., Liu, Y.-C., Kira, Z., Wang, Y.-C. F. & Huang, J.-B. (2019). A Closer Look at Few-shot Classification. *International Conference on Learning Representations (ICLR)*.

Chen, X., Hsieh, C.-J. & Gong, B. (2022). When vision transformers outperform ResNets without pre-training or strong data augmentations.

Chen, X., Dai, H., Li, Y., Gao, X. & Song, L. (2020a). Learning to stop while learning to predict. *International Conference on Machine Learning (ICML)*.

Chen, Y., Zhu, X., Li, W. & Gong, S. (2020b). Semi-supervised learning under class distribution mismatch. *Conference on Artificial Intelligence (AAAI)*.

Chernoff, H. et al. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4), 493–507.

Cichocki, A. & Amari, S.-I. (2010). Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6), 1532–1568.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. 26, 2292–2300.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dengyong, Z., Bousquet, O., Lal, T. N., Weston, J. & Schölkopf, B. (2004). Learning with local and global consistency. *Advances in Neural Information Processing Systems (NeurIPS)*.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer.

Dhillon, G. S., Chaudhari, P., Ravichandran, A. & Soatto, S. (2020). A baseline for few-shot image classification. *International Conference on Learning Representations (ICLR)*.

Dong, N. & Xing, E. (2018). Few-Shot Semantic Segmentation with Prototype Learning. *British Machine Vision Conference (BMVC)*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J. & Zisserman, A. (2010). The pascal visual object classes (voc) challenge.

Fei-Fei, L., Fergus, R. & Perona, P. (2006). One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence (PAMI)*.

Finn, C., Xu, K. & Levine, S. (2018). Probabilistic model-agnostic meta-learning. *Advances in Neural Information Processing Systems (NeurIPS)*.

Finn, C. et al. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *International Conference on Machine Learning (ICML)*.

Gairola, S., Hemani, M., Chopra, A. & Krishnamurthy, B. (2020). SimPropNet: Improved Similarity Propagation for Few-shot Image Segmentation. *International Joint Conference on Artificial Intelligence (IJCAI)*.

Ge, Z., Demyanov, S., Chen, Z. & Garnavi, R. (2017). Generative openmax for multi-class open set classification.

Gidaris, S., Bursuc, A., Komodakis, N., Pérez, P. & Cord, M. (2019). Boosting few-shot visual learning with self-supervision. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Goldblum, M., Reich, S., Fowl, L., Ni, R., Cherepanova, V. & Goldstein, T. (2020). Unraveling meta-learning: Understanding feature representations for few-shot tasks. *International Conference on Machine Learning (ICML)*.

Gordon, J., Bronskill, J., Bauer, M., Nowozin, S. & Turner, R. (2019). Meta-Learning Probabilistic Inference for Prediction. *International Conference on Learning Representations (ICLR).*

Grandvalet, Y. & Bengio, Y. (2005). Semi-supervised learning by entropy minimization. *Advances in neural information processing systems (NeurIPS).*

Guo, Y. & Cheung, N.-M. (2020). Attentive Weights Generation for Few Shot Learning via Information Maximization. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Guo, Y., Codella, N., Karlinsky, L., Codella, J. V., Smith, J. R., Saenko, K., Rosing, T. & Feris, R. (2020). A Broader Study of Cross-Domain Few-Shot Learning. *European Conference on Computer Vision (ECCV).*

Havrda, J. & Charvát, F. (1967). Quantification method of classification processes – Concept of structural a-entropy. *Kybernetika*, 3(1), 30–35.

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 770–778.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A. & Bengio, Y. (2019). Learning deep representations by mutual information estimation and maximization. *International Conference on Learning Representations (ICLR).*

Hou, R., Chang, H., Bingpeng, M., Shan, S. & Chen, X. (2019). Cross Attention Network for Few-shot Classification. *Advances in Neural Information Processing Systems (NeurIPS).*

Hu, S. X., Moreno, P. G., Xiao, Y., Shen, X., Obozinski, G., Lawrence, N. D. & Damianou, A. (2020). Empirical Bayes Transductive Meta-Learning with Synthetic Gradients. *International Conference on Learning Representations (ICLR).*

Hu, W., Miyato, T., Tokui, S., Matsumoto, E. & Sugiyama, M. (2017). Learning Discrete Representations via Information Maximizing Self-Augmented Training. *International Conference on Machine Learning (ICML).*

Hu, Y., Gripon, V. & Pateux, S. (2021). Leveraging the feature distribution in transfer-based few-shot learning. *International Conference on Artificial Neural Networks.*

Huang, S., Ma, J., Han, G. & Chang, S.-F. (2022). Task-Adaptive Negative Envision for Few-Shot Open-Set Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7171–7180.

Jabi, M., Pedersoli, M., Mitiche, A. & Ayed, I. B. (2019). Deep clustering: On the link between discriminative models and k-means. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*.

Jeong, M., Choi, S. & Kim, C. (2021). Few-shot open-set recognition by transformation consistency. *Computer Vision and Pattern Recognition Conference (CVPR)*.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z. & Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. *International Conference on Machine Learning*, pp. 4904–4916.

Joachims, T. (1999). Transductive Inference for Text Classification Using Support Vector Machines. *Proceedings of the Sixteenth International Conference on Machine Learning (ICML)*.

Killamsetty, K., Zhao, X., Chen, F. & Iyer, R. (2021). RETRIEVE: Coreset Selection for Efficient and Robust Semi-Supervised Learning.

Kim, J., Kim, T., Kim, S. & Yoo, C. D. (2019). Edge-labeling graph neural network for few-shot learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kingma, D. P. & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)*.

Krause, A., Perona, P. & Gomes, R. G. (2010). Discriminative clustering by regularized information maximization. *Advances in neural information processing systems (NeurIPS)*.

Laenen, S. & Bertinetto, L. (2021). On episodes, prototypical networks, and few-shot learning. *Advances in Neural Information Processing Systems*, 34, 24581–24592.

Lazarou, M., Stathaki, T. & Avrithis, Y. (2021). Iterative label cleaning for transductive and semi-supervised few-shot learning. *ICCV*.

Lee, H. B., Lee, H., Na, D., Kim, S., Park, M., Yang, E. & Hwang, S. J. (2019a). Learning to balance: Bayesian meta-learning for imbalanced and out-of-distribution tasks.

Lee, K., Maji, S., Ravichandran, A. & Soatto, S. (2019b). Meta-learning with differentiable convex optimization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, X., Wei, T., Chen, Y. P., Tai, Y.-W. & Tang, C.-K. (2020). FSS-1000: A 1000-class dataset for few-shot segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Liang, J., Hu, D. & Feng, J. (2020). Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation. *International Conference on Machine Learning (ICML).*

Lichtenstein, M., Sattigeri, P., Feris, R., Giryes, R. & Karlinsky, L. (2020). Tafssl: Task-adaptive feature sub-space learning for few-shot classification. *European Conference on Computer Vision (ECCV).*

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. *European conference on computer vision (ECCV).*

Linsker, R. (1988). Self-organization in a perceptual network. *Computer.*

Liu, B., Cao, Y., Lin, Y., Li, Q., Zhang, Z., Long, M. & Hu, H. (2020a). Negative Margin Matters: Understanding Margin in Few-shot Classification. *European Conference on Computer Vision (ECCV).*

Liu, B., Kang, H., Li, H., Hua, G. & Vasconcelos, N. (2020b). Few-shot open-set recognition using meta-learning. *Computer Vision and Pattern Recognition Conference (CVPR).*

Liu, J., Song, L. & Qin, Y. (2019a). Prototype Rectification for Few-Shot Learning. *arXiv preprint arXiv:1911.10713.*

Liu, J., Song, L. & Qin, Y. (2020c). Prototype rectification for few-shot learning. *European Conference on Computer Vision (ECCV).*

Liu, L., Zhou, T., Long, G., Jiang, J., Yao, L. & Zhang, C. (2019b). Prototype Propagation Networks (PPN) for Weakly-supervised Few-shot Learning on Category Graph.

Liu, W., Zhang, C., Lin, G. & Liu, F. (2020d). CRNet: Cross-Reference Networks for Few-Shot Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Liu, Y., Lee, J., Park, M., Kim, S., Yang, E., Hwang, S. J. & Yang, Y. (2019c). Learning to propagate labels: Transductive propagation network for few-shot learning. *International Conference on Learning Representations (ICLR).*

Liu, Y., Schiele, B. & Sun, Q. (2020e). An ensemble of epoch-wise empirical bayes for few-shot learning. *European Conference on Computer Vision (ECCV).*

Liu, Y., Zhang, X., Zhang, S. & He, X. (2020f). Part-aware Prototype Network for Few-shot Semantic Segmentation. *European Conference on Computer Vision (ECCV).*

Maji, S., Rahtu, E., Kannala, J., Blaschko, M. & Vedaldi, A. (2013). Fine-grained visual classification of aircraft.

Miller, E., Matsakis, N. & Viola, P. (2000a). Learning from One Example through Shared Densities on Transforms. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Miller, E. G., Matsakis, N. E. & Viola, P. A. (2000b). Learning from one example through shared densities on transforms. *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Mishra, N., Rohaninejad, M., Chen, X. & Abbeel, P. A. (2018). A simple neural attentive meta-learner. *International Conference on Learning Representations (ICLR).*

Miyato, T., Maeda, S.-i., Koyama, M. & Ishii, S. (2018). Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence (PAMI).*

Neal, L., Olson, M., Fern, X., Wong, W.-K. & Li, F. (2018). Open set learning with counterfactual images. *European Conference on Computer Vision (ECCV).*

Nguyen, K. & Todorovic, S. (2019). Feature weighting and boosting for few-shot segmentation. *International Conference on Computer Vision (ICCV).*

Nichol, A., Achiam, J. & Schulman, J. (2018). On First-Order Meta-Learning Algorithms.

Ochal, M., Patacchiola, M., Storkey, A., Vazquez, J. & Wang, S. (2021). Few-Shot Learning with Class Imbalance.

Oord, A. v. d., Li, Y. & Vinyals, O. (2018). Representation learning with contrastive predictive coding.

Oreshkin, B., López, P. R. & Lacoste, A. (2018). Tadam: Task dependent adaptive metric for improved few-shot learning. *Neural Information Processing Systems (NeurIPS).*

Qiao, L., Shi, Y., Li, J., Wang, Y., Huang, T. & Tian, Y. (2019). Transductive Episodic-Wise Adaptive Metric for Few-Shot Learning. *Proceedings of the IEEE International Conference on Computer Vision (ICCV).*

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, pp. 8748–8763.

Rakelly, K., Shelhamer, E., Darrell, T., Efros, A. & Levine, S. (2018). Conditional networks for few-shot semantic segmentation. *International Conference on Learning Representations (ICLR) Workshop.*

Ramaswamy, S., Rastogi, R. & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. *International Conference on Management of Data.*

Ravi, S. & Larochelle, H. (2016). Optimization as a model for few-shot learning. *International Conference on Learning Representations (ICLR).*

Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J. B., Larochelle, H. & Zemel, R. S. (2018). Meta-learning for semi-supervised few-shot classification. *International Conference on Learning Representations (ICLR).*

Ridnik, T., Ben-Baruch, E., Noy, A. & Zelnik-Manor, L. (2021). Imagenet-21k pretraining for the masses.

Rusu, A. A., Rao, D., Sygnowski, J., Vinyals, O., Pascanu, R., Osindero, S. & Hadsell, R. (2019). Meta-learning with latent embedding optimization. *International Conference on Learning Representations (ICLR).*

Saito, K., Kim, D. & Saenko, K. (2021). OpenMatch: Open-Set Semi-supervised Learning with Open-set Consistency Regularization.

Scheirer, W. J., de Rezende Rocha, A., Sapkota, A. & Boult, T. E. (2012). Toward open set recognition.

Schmidhuber, J. (1987). *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook.* (Ph.D. thesis, Technische Universität München).

Schroeder, B. & Cui, Y. (2018). Fgvcx fungi classification challenge 2018. Retrieved from: github.com/visipedia/fgvcx_fungi_comp.

Shaban, A., Bansal, S., Liu, Z., Essa, I. & Boots, B. (2018). One-shot learning for semantic segmentation. *British Machine Vision Conference (BMVC).*

Siam, M., Oreshkin, B. N. & Jagersand, M. (2019). AMP: Adaptive masked proxies for few-shot segmentation. *International Conference on Computer Vision (ICCV*.

Snell, J., Swersky, K. & Zemel, R. S. (2017). Prototypical networks for few-shot learning.

Sun, Q., Liu, Y., Chua, T.-S. & Schiele, B. (2019). Meta-transfer learning for few-shot learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H. & Hospedales, T. M. (2018, 06). Learning to Compare: Relation Network for Few-Shot Learning. *The IEEE Conference on Computer Vision and Pattern Recognition (Computer Vision and Pattern Recognition Conference (CVPR))*.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B. & Schmidt, L. (2020). Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33, 18583–18599.

Tian, Y., Wang, Y., Krishnan, D., Tenenbaum, J. B. & Isola, P. (2020a). Rethinking few-shot image classification: a good embedding is all you need? *European Conference on Computer Vision (ECCV)*.

Tian, Z., Zhao, H., Shu, M., Yang, Z., Li, R. & Jia, J. (2020b). Prior Guided Feature Enrichment Network for Few-Shot Segmentation.

Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J. et al. (2021). Mlp-mixer: An all-mlp architecture for vision.

Triantafillou, E., Zhu, T., Dumoulin, V., Lamblin, P., Evci, U., Xu, K., Goroshin, R., Gelada, C., Swersky, K., Manzagol, P.-A. et al. (2020). Meta-dataset: A dataset of datasets for learning to learn from few examples.

Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *Journal of statistical physics*, 52(1), 479–487.

Tseng, H.-Y., Lee, H.-Y., Huang, J.-B. & Yang, M.-H. (2020). Cross-domain few-shot classification via learned feature-wise transformation. *International Conference on Learning Representations (ICLR)*.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks*.

Vaze, S., Han, K., Vedaldi, A. & Zisserman, A. (2022). Open-Set Recognition: A Good Closed-Set Classifier is All You Need. *International Conference on Learning Representations (ICLR)*.

Veilleux, O., Boudiaf, M., Piantanida, P. & Ben Ayed, I. (2021). Realistic evaluation of transductive few-shot learning.

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K. & Wierstra, D. (2016a). *Neural Information Processing Systems (NeurIPS)*.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D. et al. (2016b). Matching networks for one shot learning.

Wang, H., Zhang, X., Hu, Y., Yang, Y., Cao, X. & Zhen, X. (2020a). Few-Shot Semantic Segmentation with Democratic Attention Networks. *ECCV*.

Wang, K., Liew, J. H., Zou, Y., Zhou, D. & Feng, J. (2019a). PANet: Few-shot image semantic segmentation with prototype alignment. *International Conference on Computer Vision (ICCV)*.

Wang, Y., Chao, W.-L., Weinberger, K. Q. & van der Maaten, L. (2019b). Simpleshot: Revisiting nearest-neighbor classification for few-shot learning.

Wang, Y., Xu, C., Liu, C., Zhang, L. & Fu, Y. (2020b). Instance credibility inference for few-shot learning. *Computer Vision and Pattern Recognition Conference (CVPR)*.

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S. & Perona, P. (2010). *Caltech-UCSD Birds 200* (Report n°CNS-TR-2010-001).

Wertheimer, D. & Hariharan, B. (2019). Few-shot learning with localization in realistic settings. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wightman, R. (2019). PyTorch Image Models. GitHub.

Yalniz, I. Z., Jégou, H., Chen, K., Paluri, M. & Mahajan, D. (2019). Billion-scale semi-supervised learning for image classification.

Yang, B., Liu, C., Li, B., Jiao, J. & Ye, Q. (2020a). Prototype Mixture Models for Few-shot Semantic Segmentation. *European Conference on Computer Vision (ECCV)*.

Yang, B., Liu, C., Li, B., Jiao, J. & Ye, Q. (2020b). Prototype mixture models for few-shot semantic segmentation. *European Conference on Computer Vision*, pp. 763–778.

Yang, L., Li, L., Zhang, Z., Zhou, X., Zhou, E. & Liu, Y. (2020c). DPGN: Distribution Propagation Graph Network for Few-shot Learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, Y., Meng, F., Li, H., Wu, Q., Xu, X. & Chen, S. (2020d). A new local transformation module for few-shot segmentation. *International Conference on Multimedia Modeling (ICMM)*.

Ye, H.-J., Hu, H., Zhan, D.-C. & Sha, F. (2020a). Learning embedding adaptation for few-shot learning. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ye, H.-J., Hu, H., Zhan, D.-C. & Sha, F. (2020b). Few-shot learning via embedding adaptation with set-to-set functions. *Computer Vision and Pattern Recognition Conference (CVPR)*.

Yu, Q., Ikami, D., Irie, G. & Aizawa, K. (2020). Multi-task curriculum framework for open-set semi-supervised learning. *European Conference on Computer Vision (ECCV)*.

Zhang, C., Lin, G., Liu, F., Guo, J., Wu, Q. & Yao, R. (2019a). Pyramid Graph Networks with Connection Attentions for Region-Based One-Shot Semantic Segmentation. *International Conference on Computer Vision (ICCV)*.

Zhang, C., Lin, G., Liu, F., Yao, R. & Shen, C. (2019b). CANet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, C., Cai, Y., Lin, G. & Shen, C. (2020a). DeepEMD: Few-Shot Image Classification With Differentiable Earth Mover's Distance and Structured Classifiers. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

Zhang, X., Wei, Y., Yang, Y. & Huang, T. S. (2020b). SG-one: Similarity guidance network for one-shot semantic segmentation.

Zhao, H., Shi, J., Qi, X., Wang, X. & Jia, J. (2017). Pyramid scene parsing network. *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Zhao, Y., Nasrullah, Z. & Li, Z. (2019). PyOD: A Python Toolbox for Scalable Outlier Detection.

Zhen, X., Sun, H., Du, Y., Xu, J., Yin, Y., Shao, L. & Snoek, C. (2020). Learning to learn kernels with variational random features. *International Conference on Machine Learning (ICML)*.

Zhou, D.-W., Ye, H.-J. & Zhan, D.-C. (2021). Learning placeholders for open-set recognition. *Computer Vision and Pattern Recognition Conference (CVPR).*

Ziko, I., Dolz, J., Granger, E. & Ayed, I. B. (2020). Laplacian regularized few-shot learning. *International Conference on Machine Learning (ICML).*